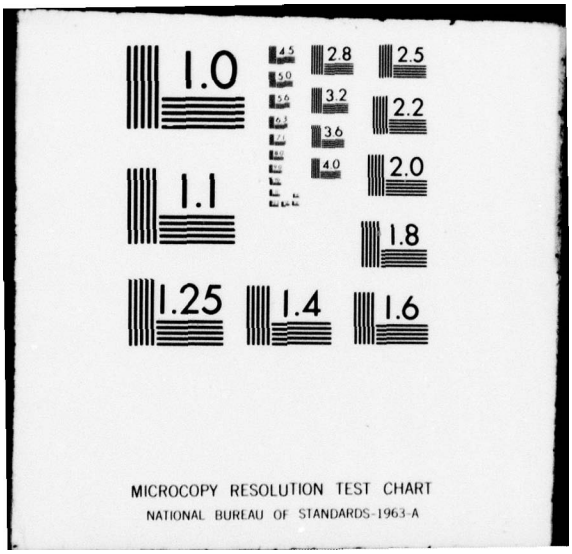
AD-A063 120 NAVAL RESEARCH LAB WASHINGTON D C F/G 9/4
AXIOMATIC DERIVATION OF THE PRINCIPLE OF MAXIMUM ENTROPY AND TH--ETC(U)
DEC 78 J E SHORE, R W JOHNSON

UNCLASSIFIED NRL-MR-3898 NL

END
DATE FILMED
3-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NRL Memorandum Report 3898	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*: AXIOMATIC DERIVATION OF THE PRINCIPLE OF MAXIMUM ENTROPY AND THE PRINCIPLE OF MINIMUM CROSS-ENTROPY	5. TYPE OF REPORT & PERIOD COVERED: Interim report on a continuing NRL problem.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): J. E. Shore and R. W. Johnson	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Research Laboratory, Washington, D.C. 20375	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NRL Problem B02-35, 61153N, RR014-09-41
11. CONTROLLING OFFICE NAME AND ADDRESS: Department of the Navy, Office of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: December 15, 1978
	13. NUMBER OF PAGES: 62
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*	15. SECURITY CLASS. *(of this report)*: UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Maximum entropy
Minimum cross- entropy
Directed divergence
Minimum discrimination information
Information theory
Inference
Pattern recognition

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

We prove that, in a well-defined sense, Jaynes's principle of maximum entropy and Kullback's principle of minimum cross-entropy (minimum directed divergence) provide uniquely correct, general methods of inductive inference when new information is given in the form of expected values. Previous justifications rely heavily on intuitive arguments and on the properties of entropy and cross-entropy as information measures. Our approach assumes that reasonable methods of inductive inference should lead to consistent results whenever there are different ways of taking the same

(Continues) next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

20. Abstract (Continued)

information into account --- for example, in different coordinate systems. We formalize this requirement as four consistency axioms stated in terms of an abstract information operator; the axioms make no reference to information measures. We then prove that the principle of maximum entropy is correct in the following sense: maximizing any other function but entropy will lead to inconsistency unless that function and entropy have identical maxima. Stated differently, we prove that, given information in the form of constraints on expected values, there is only one distribution satisfying the constraints that can be chosen as the result of a procedure that satisfies the consistency axioms; this unique distribution can be obtained by maximizing entropy. We establish this result both directly and as a special case (uniform priors) of an analogous, more general result for the principle of minimum cross-entropy. We obtain results both for continuous probability densities and for discrete distributions.

Contents

I. INTRODUCTION

The purpose of this paper is to prove that Jaynes's principle of maximum entropy and Kullback's principle of minimum cross-entropy (minimum directed divergence) provide correct, general methods of inductive inference when given new information in the form of expected values. Unlike previous justifications, ours does not rely on intuitive arguments or on the properties of entropy and cross-entropy as information measures.

A. The Maximum Entropy Principle and the Minimum Cross-entropy Principle

Suppose you know that a system has a finite set of possible states x_i with unknown probabilities $q^\dagger(x_i)$. Suppose you then learn the values of certain expectations $\sum_i q^\dagger(x_i)f_k(x_i)$, or bounds on these values, and you need to choose a distribution q that is in some sense the best estimate of $q^\dagger$ given what you know. In such problems, the known expectations are referred to as constraints, and distributions with expected values that equal the known values or fall within the known bounds are said to satisfy the constraints. Usually, although the constraints rule out an infinite set of distributions, there remains an infinite set of distributions that satisfy the constraints. Which one should you choose?

The principle of maximum entropy is a prescription for solving such problems. It states that, of all the distributions q that satisfy the constraints, you should choose the one with the largest entropy $-\sum_i q(x_i)\log(q(x_i))$. Entropy maximization was first proposed as a general inference procedure by Edwin Jaynes more than twenty years ago [1]. Since then, it has been applied successfully in a remarkable variety of fields, including statistical mechanics and thermodynamics [1]-[8], statistics

Note: Manuscript submitted October 23, 1978.

[9]-[11, Chapter 6], reliability estimation [11, Chapter 10], [12], traffic networks [13], queuing theory and computer system modeling [14], [15], system simulation [16], production line decision making [17], [18], computer memory reference patterns [19], system modularity [20], group behavior [21], stock market analysis [22], and general probabilistic problem solving [11], [17], [23]-[25]. Among geophysicists and radio astronomers, there is much current interest in maximum entropy spectral analysis [26]-[29].

The principle of minimum cross-entropy is a generalization that applies in cases when a prior distribution that estimates $q^{\dagger}$ is known in addition to the newly learned expectations. The principle states that, of all the distributions q that satisfy the constraints, you should choose the one with the smallest cross-entropy $\sum_i q(x_i)\log(q(x_i)/p(x_i))$, where p is the prior estimate. Minimizing cross-entropy is equivalent to maximizing entropy in cases where the prior is a uniform distribution. Unlike entropy maximization, cross-entropy minimization generalizes correctly for continuous probability densities. In this case, one minimizes the functional

$$\int dx\, q(x)\log(q(x)/p(x)) \quad . \tag{1}$$

Cross-entropy goes by other names, including expected weight of evidence [30, p. 72], directed divergence [31, p. 6], and relative entropy [32]. We prefer the term cross-entropy, which is due to Good [9]. The principle of minimum cross-entropy was first proposed by Kullback, who called it a principle of minimum directed divergence or minimum discrimination information [32, p. 37]. It has been advocated in various forms by others [9], [33], [34], including Jaynes [3], [25], who showed that generalizing entropy maximization to continuous densities leads to (1) with p(x) being called an

"invariant measure" instead of a prior density. Since entropy maximization does not deal with prior densities --- there being an implicit assumption of uniform priors --- this just expresses the fact that a uniform prior in one coordinate system may not be uniform in another. Cross-entropy minimization has been applied primarily to statistics [9], [31], [35], but also to statistical mechanics [8], chemistry [36], pattern recognition [37], [38], and the computer storage of probability distributions [39].

As a historical note, we point out that entropy maximization and cross-entropy minimization both have roots in Shannon's work [40], [41]. For discrete, noiseless systems, maximizing the source entropy results in the best source encoding, in the sense of enabling the highest information rate over a fixed capacity channel [40]. For continuous systems, Shannon's definition of source rate for a fixed fidelity criterion involved the minimization of a functional like cross-entropy [41].

The mathematics of minimizing cross-entropy subject to constraints is discussed in Appendix B.

B. Justifying the Principles as General Methods of Inference

Despite its success, the maximum entropy principle remains controversial [32], [42]-[46]. The controversy stems from what some perceive to be weaknesses in the foundations of the principle, which is usually justified on the basis of entropy's unique properties as a measure of the uncertainty represented by a probability distribution. That entropy has such unique properties is generally undisputed because one can prove, to within the choice of logarithmic base, that entropy is the only function satisfying various axioms that are accepted as requirements for an uncertainty measure [40], [47]. Intuitively, the maximum entropy principle follows quite naturally from

such axiomatic characterizations. In proposing it, Jaynes described the maximum entropy distribution as "the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have....The maximum entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information" [1, p. 623]. Elsewhere, he states that the maximum entropy distribution "agrees with what is known, but expresses 'maximum uncertainty' with respect to all other matters, and thus leaves a maximum possible freedom for our final decisions to be influenced by the subsequent sample data" [25, p. 231]. Somewhat whimsically, Benes justified his use of entropy maximization as "a reasonable and systematic way of throwing up our hands" [13, p. 234]. Others argue similarly [5]-[9], [11].

Although most of the justification for the maximum entropy principle rests on entropy's properties as an information measure, other kinds of arguments also support the principle. In response to a common objection that the maximum entropy distribution has no frequency interpretation (e.g., [42]), Jaynes showed that this distribution is equal to the frequency distribution that can be realized in the greatest number of ways [25]. He also showed that entropy maximization is consistent with various other principles of probability theory [25].

Similar justifications can be advanced for the principle of cross-entropy minimization. Like entropy, cross-entropy can be characterized axiomatically, both in the discrete case [8], [48]-[51] and in the continuous case [34]. Cross-entropy has various properties that are desirable for an information measure [33], [34], and it can be argued [48] that cross-entropy measures the

amount of information necessary to change a prior p into the posterior q. The principle of cross-entropy minimization then follows intuitively much like entropy maximization.

To some, entropy's properties as an information measure make it obvious that entropy maximization is the correct way to account for constraint information. To others, such an informal and intuitive justification yields plausibility for the maximum entropy principle, but not proof --- why maximize entropy; why not some other function?

Such questions are not answered unequivocally by previous justifications because these justifications argue indirectly --- they are based on a formal description of what is required of an information measure rather than on a formal description of what is required of a method for taking new information into account. Since the maximum entropy principle is asserted as a general method of inductive inference, it seems reasonable to require that, if there are different ways to take the same information into account, these different ways should lead to consistent results. Our approach is to formalize this requirement as a set of consistency axioms. The axioms are stated in terms of an abstract information operator; they make no reference to information measures or to properties of information measures.

We can then prove that the maximum entropy principle is correct in the following sense: maximizing any other function but entropy will lead to logical inconsistencies unless that function and entropy have identical maxima (any monotonic function of entropy will work, for example). Stated differently, we prove that, given new information in the form of constraints on expected values, there is only one distribution satisfying these constraints that can be chosen as the result of a procedure that satisfies the

consistency axioms; this unique distribution can be obtained by maximizing entropy. We establish this result both directly and as a special case of an analogous result for the principle of minimum cross entropy: We prove, for the continuous case, that minimizing any other functional but cross-entropy will lead to logical inconsistencies unless that functional and cross-entropy have identical minima. Stated differently, we prove that, given a prior density and new information in the form of constraints on expected values, there is only one posterior density satisfying these constraints that can be chosen in a manner that satisfies the axioms; this unique posterior can be obtained by minimizing cross-entropy.

We require only four axioms. Informally, they may be phrased as follows:

1) Uniqueness. The result should be unique.
2) Invariance. It shouldn't matter in which coordinate system one accounts for new information.
3) System independence. It shouldn't matter whether one accounts for independent information about independent systems separately in terms of different densities or together in terms of a joint density.
4) Subset independence. It shouldn't matter whether one accounts for information about an independent subset of system states in terms of a separate conditional density or in terms of the full system density.

All four of these axioms are based on a single fundamental principle: If a problem can be solved in more than one way, the results should be consistent.

Our approach is analogous to work of Cox [52], [53], [11, Chap. 1] and similar work of Janossy [54], [55]. They assumed that probability theory must provide a consistent model of inductive inference, and they showed how this

requirement leads to functional equations whose solutions include the standard equations of probability theory.

C. Outline

The remainder of the paper is organized as follows: In Section II we introduce some definitions and notation. In Section III we motivate the specific axioms we use and we give their formal statements. The conquences of the axioms for the general case of continuous densities are explored in Section IV in terms of a series of theorems that culminates in our main result justifying the principle of cross-entropy minimization. The discrete case, including the principle of maximum entropy, is discussed in Section V. Section VI contains a discussion of the difference between axioms of inference methods and axioms of information measures. We conclude with a brief summary in Section VII.

II. DEFINITIONS AND NOTATION

Because we need to formalize inference about probability densities that must satisfy an arbitrary set of expected value constraints, we need a concise notation to describe such arbitrary constraints and to refer to the densities that satisfy them. For these purposes it is convenient to speak in terms of sets of probability densities and to use set theory notation. We also need a concise notation for the inference procedure that minimizes some functional in order to choose a posterior density. This notation must permit us to state required properties of the inference procedure rather than required properties of the functional. We therefore introduce an abstract information operator that yields a posterior density from a prior density and new constraint information. We are then able to state inference requirements in terms of axioms for this information operator.

We use lower-case boldface Roman letters to denote system states, which may be multidimensional, and upper-case boldface Roman letters to denote sets of possible system states. We use lower-case Roman letters to denote probability densities, and upper case script letters to denote sets of probability densities. Thus, let $\underset{\sim}{x}$ denote a single state of some system that has a set $\underset{\sim}{D}$ of possible system states and a probability density $q^\dagger(\underset{\sim}{x})$ of states. Let $\mathcal{Q}$ be the set of all probability densities q on $\underset{\sim}{D}$ such that $q(\underset{\sim}{x}) \geqslant 0$ for $\underset{\sim}{x} \in \underset{\sim}{D}$ and

$$\int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q(\underset{\sim}{x}) = 1 \quad . \tag{2}$$

We assume that the existence of $q^\dagger \in \mathcal{Q}$ is known but that $q^\dagger$ itself is unknown. The density $q^\dagger$ is sometimes known as a "true" density; we use daggers $\dagger$ to indicate such densities. When we refer to a set of values $q(\underset{\sim}{x})$ for $\underset{\sim}{x} \in \underset{\sim}{S}$,

where $\underset{\sim}{S} \subseteq \underset{\sim}{D}$ is some subset of system states, we sometimes write $q(\underset{\sim}{x} \in \underset{\sim}{S})$.

We are concerned with problems in which one gains new information about the system in the form of some combination of linear equality constraints

$$\int_{\underset{\sim}{D}} d\underset{\sim}{x}\ q^{\dagger}(\underset{\sim}{x}) a_k(\underset{\sim}{x}) = 0 \tag{3}$$

and inequality constraints

$$\int_{\underset{\sim}{D}} d\underset{\sim}{x}\ q^{\dagger}(\underset{\sim}{x}) c_k(\underset{\sim}{x}) \geqslant 0 \tag{4}$$

for known sets of bounded functions a_k and c_k. The set of probability densities that satisfy such linear constraints always comprises a closed, convex subset of $\mathcal{Q}$. (A density set $\mathcal{J}$ is convex if and only if, given $0 \leq A \leq 1$ and any $q, r \in \mathcal{J}$, it contains the weighted average $Aq + (1-A)r$. Informally, $\mathcal{J}$ can be thought of as containing all possible "compromises" between q and r.) Furthermore, any closed, convex subset of $\mathcal{Q}$ can be defined by a suitable combination of equality and inequality constraints, possibly infinite in number. We are therefore concerned with problems in which the new information locates $q^{\dagger}$ to within a specified closed, convex subset of $\mathcal{Q}$. For convenience, we express constraints in these terms, using the notation $I = (q^{\dagger} \in \mathcal{J})$ to mean that $q^{\dagger}$ is a member of the closed, convex set $\mathcal{J} \subseteq \mathcal{Q}$. (Note that $\mathcal{Q}$ itself is convex.) We refer to I as a constraint and to $\mathcal{J}$ as a constraint set. We use upper case Roman letters to denote constraints.

Let $p \in \mathcal{Q}$ be some prior density that is an estimate of $q^{\dagger}$ obtained, by any means, prior to learning I. We require that priors be strictly positive:

$$p(\underset{\sim}{x} \in \underset{\sim}{D}) > 0 \quad . \tag{5}$$

(This restriction is discussed below.) Given a prior p and new information I, the posterior density $q \in \mathcal{J}$ that results from taking I into account is chosen

by minimizing the functional $H(q,p)$ in the constraint set $\mathcal{J}$. That is, the posterior q satisfies

$$H(q,p) = \min_{q' \in \mathcal{J}} H(q',p) \quad . \tag{6}$$

For convenience, we introduce an "information operator" $\circ$ that expresses (6) using the notation

$$q = p \circ I \quad . \tag{7}$$

The operator $\circ$ takes two arguments --- a prior and new information --- and yields a posterior. For some other functional $F(q,p)$, suppose q satisfies (6) if and only if it satisfies

$$F(q,p) = \min_{q' \in \mathcal{J}} F(q',p) \quad .$$

Then we say that F and H are <u>equivalent</u>. If F and H are equivalent, the operator $\circ$ can be realized using either functional.

If H has the form

$$H(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q(\underset{\sim}{x}) \log(q(\underset{\sim}{x})/p(\underset{\sim}{x})),$$

then (7) expresses the principle of minimum cross-entropy. At this point, however, we assume only that H is well-behaved. In Section III, we give consistency axioms for the operator $\circ$ that restrict the form of H in ways we investigate in Section IV. We say that a functional H <u>satisfies</u> one of these axioms if the axiom is satisfied by the operator $\circ$ that is realized using H.

The restriction (5) to strictly positive priors reflects our assumption that $p(\underset{\sim}{x}) = 0$ would indicate the impossibility of $\underset{\sim}{x}$, whereas we assume that $\underset{\sim}{D}$ is the set of possible states in the sense that prior information has not

ruled out any state $\underset{\sim}{x} \in \underset{\sim}{D}$. We do not impose a similar restriction on the posterior $q = p \circ I$ since the new information I may render impossible states currently thought to be possible. If this happens, then $\underset{\sim}{D}$ must be redefined before q is used as a prior in some further application of the operator $\circ$. The restriction (5) is not necessary for our results, and it does not restrict them in any significant way, but it does help in avoiding certain technical problems that would otherwise result from division by $p(\underset{\sim}{x})$.

For some subset $\underset{\sim}{S} \subseteq \underset{\sim}{D}$ of system states and $\underset{\sim}{x} \in \underset{\sim}{S}$, let

$$q(\underset{\sim}{x} \mid \underset{\sim}{x} \in \underset{\sim}{S}) = q(\underset{\sim}{x}) \Big/ \int_{\underset{\sim}{S}} d\underset{\sim}{x}' \; q(\underset{\sim}{x}') \tag{8}$$

be the <u>conditional density</u>, given $\underset{\sim}{x} \in \underset{\sim}{S}$, corresponding to any $q \in \mathbb{Q}$. We use the equation

$$q(\underset{\sim}{x} \mid \underset{\sim}{x} \in \underset{\sim}{S}) = q^{*}\underset{\sim}{S} \tag{9}$$

as a shorthand notation for (8).

In cases where $\underset{\sim}{D}$ is a discrete set of system states, densities are replaced by discrete distributions and integrals by sums in the usual way. We use lower-case boldface roman letters to denote discrete probability distributions, which we consider to be vectors, for example $\underset{\sim}{q} = q_1, q_2, \ldots, q_n$. This results in some potential confusion --- for example, the symbol $\underset{\sim}{r}$ could refer to a system state or a discrete distribution, and s_i could refer to a probability density or a component of a discrete distribution --- but the intended meaning is always made clear by the context.

III. THE AXIOMS

We precede the formal statement of each axiom with a justification. We assume, throughout, a system with possible states $\underset{\sim}{D}$ and probability density $q^\dagger \in \mathcal{Q}$.

A. Uniqueness

If we solve the same problem twice in exactly the same way, we expect the same answer to result in both cases. Stated differently, if $p_1 = p_2$ holds, we want $p_1 \cdot I = p_2 \cdot I$ to hold as well. Such consistency cannot be expected unless the following axiom holds:

> Axiom I (uniqueness): The posterior $q = p \cdot I$ is unique for any prior $p \in \mathcal{Q}$ and new information $I = (q^\dagger \in \mathcal{J})$, where $\mathcal{J} \subseteq \mathcal{Q}$.

Actually, Axiom I is implicit in our notation.

B. Invariance

Similarly, we expect the same answer to result from solving the same problem in two different coordinate systems, in the sense that the posterior in one system should be the coordinate transformation of the posterior in the other system. We state this requirement formally as follows:

> Axiom II (invariance): Let Γ be a coordinate transformation from $\underset{\sim}{x} \in \underset{\sim}{D}$ to $\underset{\sim}{y} \in \Gamma\underset{\sim}{D}$ with $(\Gamma q)(\underset{\sim}{y}) = \underset{\sim}{J}^{-1} q(\underset{\sim}{x})$, where $\underset{\sim}{J}$ is the Jacobian $\underset{\sim}{J} = \partial(\underset{\sim}{y})/\partial(\underset{\sim}{x})$. Let $\Gamma\mathcal{Q}$ be the set of densities Γq corresponding to densities $q \in \mathcal{Q}$. Let $(\Gamma\mathcal{J}) \subseteq (\Gamma\mathcal{Q})$ correspond to $\mathcal{J} \subseteq \mathcal{Q}$. Then, for any prior $p \in \mathcal{Q}$ and new information $I = (q^\dagger \in \mathcal{J})$,
>
> $$(\Gamma p) \cdot (\Gamma I) = \Gamma(p \cdot I) \qquad (10)$$
>
> holds, where $\Gamma I = ((\Gamma q^\dagger) \in (\Gamma\mathcal{J}))$.

C. System Independence

Suppose there are two systems, one with a set $\underset{\sim}{D}_1$ of system states and probability density of states $q_1^\dagger \in \mathcal{Q}_1$, and the other with a set $\underset{\sim}{D}_2$ of system states and probability density of states $q_2^\dagger \in \mathcal{Q}_2$. We also describe the two systems jointly using the joint probability density $q^\dagger(\underset{\sim}{x}_1, \underset{\sim}{x}_2)$, where $\underset{\sim}{x}_1 \in \underset{\sim}{D}_1$, $\underset{\sim}{x}_2 \in \underset{\sim}{D}_2$, and $q \in \mathcal{Q}_{12}$. If the two systems were independent, then the joint density would satisfy

$$q^\dagger(\underset{\sim}{x}_1, \underset{\sim}{x}_2) = q_1^\dagger(\underset{\sim}{x}_1) q_2^\dagger(\underset{\sim}{x}_2). \tag{11}$$

Now suppose that we have prior densities p_1 and p_2 for the two systems, and suppose that we obtain separate new information $I_1 = (q_1^\dagger \in \mathcal{J}_1)$ about one system and $I_2 = (q_2^\dagger \in \mathcal{J}_2)$, where $\mathcal{J}_1 \subseteq \mathcal{Q}_1$ and $\mathcal{J}_2 \subseteq \mathcal{Q}_2$. Such new information can also be expressed completely in terms of the joint density $q^\dagger$. For example I_1 can be expressed as $I_1 = (q^\dagger \in \mathcal{J}_1')$, where $\mathcal{J}_1' \subseteq \mathcal{Q}_{12}$ is the set of joint densities $q \in \mathcal{Q}_{12}$ such that $q_1 \in \mathcal{J}_1$, where

$$q_1(\underset{\sim}{x}_1) = \int_{\underset{\sim}{D}_2} d\underset{\sim}{x}_2 \; q(\underset{\sim}{x}_1, \underset{\sim}{x}_2) \; .$$

I_2 can be expressed similarly in terms of the joint density $q^\dagger$ instead of in terms of $q_2^\dagger$. Now, since the two priors together define a joint prior $p_{12} = p_1 p_2$, it follows that there are two ways to take the new information I_1 and I_2 into account: We can obtain separate posteriors $q_1 = p_1 \circ I_1$ and $q_2 = p_2 \circ I_2$, or we can obtain a joint posterior $q = p_{12} \circ (I_1 \wedge I_2)$. Because p_1 and p_2 are independent, and because I_1 and I_2 give no information about any interaction between the two systems, we expect these two ways to be related by $q_{12} = q_1 q_2$, whether

or not (11) in fact holds. We therefore have the following axiom:

Axiom III (system independence): Let there be two systems, one with a set $\underset{\sim}{D}_1$ of system states and probability density of states $q_1^\dagger \in \mathcal{Q}_1$, and the other with a set $\underset{\sim}{D}_2$ of system states and probability density of states $q_2^\dagger \in \mathcal{Q}_2$. Let $p_1 \in \mathcal{Q}_1$ and $p_2 \in \mathcal{Q}_2$ be prior densities for the two systems. Let $I_1 = (q_1^\dagger \in \mathcal{J}_1)$ and $I_2 = (q_2^\dagger \in \mathcal{J}_2)$, be new information about the two systems, where $\mathcal{J}_1 \subseteq \mathcal{D}_1$ and $\mathcal{J}_2 \subseteq \mathcal{D}_2$.

Then (12)

$$(p_1 p_2) \circ (I_1 \wedge I_2) = (p_1 \circ I_1)(p_2 \circ I_2)$$

holds.

D. Subset Independence

Our final axiom concerns situations in which the set of system states $\underset{\sim}{D}$ decomposes naturally into a number of disjoint subsets $\underset{\sim}{S}_1$, $\underset{\sim}{S}_2$, ..., $\underset{\sim}{S}_n$ whose union is $\underset{\sim}{D}$. As usual, we assume a known prior $p \in \mathcal{Q}$. Suppose, for each subset $\underset{\sim}{S}_i$, we obtain new information about the conditional density $q^\dagger * \underset{\sim}{S}_i$, namely $I_i = (q^\dagger * \underset{\sim}{S}_i \in \mathcal{J}_i)$, where $\mathcal{J}_i \subseteq \mathcal{S}_i$ and $\mathcal{S}_i$ is the set of conditional densities on $\underset{\sim}{S}_i$ (see (8)-(9)). One way of accounting for this information is to obtain, for each conditional density, a conditional posterior $q_i = (p * \underset{\sim}{S}_i) \circ I_i$ from the conditional prior $p * \underset{\sim}{S}_i$. Another way is to obtain a posterior $q = p \circ I$ for the whole system, where $I = I_1 \wedge I_2 \wedge \ldots \wedge I_n$. We expect that the conditional density $q * \underset{\sim}{S}_i$ be the same as the conditional density q_i obtained the previous way. That is, we expect (13)

$$(p \circ I) * \underset{\sim}{S}_i = (p * \underset{\sim}{S}_i) \circ I_i$$

to hold.

Moreover, suppose that we also learn the probability of being in each of the n subsets. That is, we learn $M = (q^\dagger \in \mathcal{M})$, where $\mathcal{M}$ is the set of densities q that satisfy

$$\int_{\underset{\sim}{S}_i} d\underset{\sim}{x}\ q(\underset{\sim}{x}) = m_i$$

for each subset $\underset{\sim}{S}_i$. The known numbers m_i are the probabilities that the system is in a state within $\underset{\sim}{S}_i$. The m_i satisfy $\sum_i m_i = 1$. Taking M into account should not affect the conditional densities that result from taking I into account. We therefore expect a more general version of (13) to hold, namely,

$$(p \circ (I \wedge M)) * \underset{\sim}{S}_i = (p * \underset{\sim}{S}_i) \circ I_i .$$

We restate this formally as our final axiom:

Axiom IV (subset independence): Let $\underset{\sim}{S}_1, \underset{\sim}{S}_2, \ldots, \underset{\sim}{S}_n$ be disjoint subsets whose union is $\underline{D}$, and let $p \in \mathcal{Q}$ be any known prior. For each subset $\underset{\sim}{S}_i$, let $I_i = (q^\dagger * \underset{\sim}{S}_i \in \mathcal{J}_i)$ be new information about the conditional density $q^\dagger * \underset{\sim}{S}_i$, where $\mathcal{J}_i \subseteq \mathcal{S}_i$ and $\mathcal{S}_i$ is the set of densities on $\underset{\sim}{S}_i$. Let $M = (q^\dagger \in \mathcal{M})$ be new information giving the probability of being in each of the n subsets, where $\mathcal{M}$ is the set of densities q that satisfy

$$\int_{\underset{\sim}{S}_i} d\underset{\sim}{x}\ q(\underset{\sim}{x}) = m_i \qquad (14)$$

for each subset $\underset{\sim}{S}_i$, where the m_i are known values. Then

$$(p \circ (I \wedge M)) * \underset{\sim}{S}_i = (p * \underset{\sim}{S}_i) \circ I_i \qquad (15)$$

holds, where $I = I_1 \wedge I_2 \wedge \ldots \wedge I_n$.

IV. CONSEQUENCES OF THE AXIOMS

A. Summary

Since we require the axioms to hold for both equality and inequality constraints (2)-(3), they must hold for equality constraints alone. We first investigate the axioms' consequences assuming only equality constraints. Later, we show that the resulting restricted form for H also satisfies the axioms in the case of inequality constraints.

We establish our main result in four steps. The first shows that the subset independence axiom and a special case of the invariance axiom together restrict H(q,p) to functionals that are equivalent to the form

$$F(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\ f(q(\underset{\sim}{x}),p(\underset{\sim}{x})) \qquad (16)$$

for some function f of two variables. This form, which we call the "sum form", is really the simplest that H could have and is the most convenient mathematically for the purpose of minimization. In the axiomatic characterizations in [34], [49], and [50], the sum form was assumed rather than derived.

Although a special case of invariance is invoked in deriving (16), the sum form in general does not satisfy the invariance axiom. Our next step is to show that general invariance restricts the possible forms of the function f so that H is in turn restricted to functionals that are equivalent to the form

$$F(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\ q(\underset{\sim}{x})h(q(\underset{\sim}{x})/p(\underset{\sim}{x})), \qquad (17)$$

where h is some function of a single variable. Our third step is to apply the system independence axiom. The result restricts the possible forms of the

function h and shows that, if H is a functional that satisfies all four axioms, then H is equivalent to the functional

$$F(q,p) = \int_{\underset{\sim}{x}} d\underset{\sim}{x}\; q(\underset{\sim}{x})\log(q(\underset{\sim}{x})/p(\underset{\sim}{x})), \qquad (18)$$

i.e, H is equivalent to cross-entropy. Since it could still be imagined that no functional satisfies the axioms, our final step is to show that cross-entropy does. We do this in the general case of equality and inequality constraints.

B. Deriving the Sum Form

We derive the sum form in several steps. First, we show that, when the assumptions of the subset independence axiom hold, the posterior values within any subspace are independent of the values in the other subspaces. Next, we move formally to the discrete case and show that invariance implies that H is equivalent to a symmetric function. We then apply the subset independence axiom and prove that H is equivalent to functions of the form $F(\underset{\sim}{q},\underset{\sim}{p}) = \sum_j f(q_j,p_j)$, where $\underset{\sim}{p} = p_1, p_2,\ldots, p_n$ and $\underset{\sim}{q} = q_1, q_2,\ldots, q_n$ are discrete prior and posterior distributions respectively, and we return to the continuous case yielding (16).

We begin with the following lemma:

Lemma I: Let the assumptions of Axiom IV hold, and let $q = p\bullet(I \wedge M)$ be the posterior for the whole system ($q \in \mathcal{Q}$). Then $q(\underset{\sim}{x} \in \underset{\sim}{S}_i)$ is functionally independent of $q(\underset{\sim}{x} \notin \underset{\sim}{S}_i)$, of the prior $p(\underset{\sim}{x} \notin \underset{\sim}{S}_i)$, and of n.

Proof: Let

$$q_i = (p^{*}\underset{\sim}{S}_i)\bullet I_i \qquad (19)$$

be the conditional posterior density in the ith subspace ($q_i \in \mathcal{S}_i$). Since

$p^*\underset{\sim}{S}_i$ depends on p only in terms of $p(\underset{\sim}{x} \in \underset{\sim}{S}_i)$ (see (8)-(9)), so does q_i. Furtherore, since q_i is the solution (19) to a problem in which $\underset{\sim}{x} \in \underset{\sim}{S}_i$ only, q_i cannot depend on $q(\underset{\sim}{x} \notin \underset{\sim}{S}_i)$. Now, (15) states that $(q(\underset{\sim}{x})/m_i) = q_i(\underset{\sim}{x})$ or

$$q(\underset{\sim}{x}) = m_i q_i(\underset{\sim}{x})$$

for $\underset{\sim}{x} \in \underset{\sim}{S}_i$, where we have used (8) and (14). Since the m_i are fixed numbers, it follows that $q(\underset{\sim}{x} \in \underset{\sim}{S}_i)$ is independent of $q(\underset{\sim}{x} \notin \underset{\sim}{S}_i)$ and $p(\underset{\sim}{x} \notin \underset{\sim}{S}_i)$. This proves Lemma I.

Our next step is to transform to the discrete case in the formal manner given in the following lemma:

Lemma II. Let $\underset{\sim}{S}_1, \underset{\sim}{S}_2, \ldots, \underset{\sim}{S}_n$ be disjoint subsets whose union is $\underset{\sim}{D}$. For a prior p and a posterior $q = p \cdot I$, let

$$p_j = \int_{\underset{\sim}{S}_j} d\underset{\sim}{x}\, p(\underset{\sim}{x})$$

and

$$q_j = \int_{\underset{\sim}{S}_j} d\underset{\sim}{x}\, q(\underset{\sim}{x}) \quad .$$

Suppose that $p(\underset{\sim}{x} \in \underset{\sim}{S}_j)$ is constant, i.e., that p takes on constant values in each subset. Furthermore, let the new information I be provided by a set of constraints (3)-(4) in which the known functions a_k and c_k are also constant in each subset. Then the posterior $q = p \cdot I$ must also be constant in each subset, and H is equivalent to a symmetric function of the n pairs of variables (q_j, p_j). We refer to this situation as the discrete case.

Proof: Since the known functions a_k and c_k are constant in each subset, the constraints have the form

$$\sum_j q_j^{\dagger} a_{kj} = 0 \qquad (20)$$

or

$$\sum_j q_j^\dagger c_{kj} \geqslant 0, \tag{21}$$

where $a_{kj} = a_k(\underset{\sim}{x} \in \underset{\sim}{S}_j)$, $c_{kj} = c_k(\underset{\sim}{x} \in \underset{\sim}{S}_j)$, and

$$q_j^\dagger = \int_{\underset{\sim}{S}_j} d\underset{\sim}{x}\ q^\dagger(\underset{\sim}{x}).$$

Now, let Γ be a measure-preserving transformation that scrambles the $\underset{\sim}{x}$ within each subset $\underset{\sim}{S}_i$. This leaves the prior unchanged and it leaves the constraints (20)-(21) unchanged. It follows from invariance (10) that Γ must also leave q unchanged, which will only be the case if q is constant in each subset $\underset{\sim}{S}_i$. With q and p each taking on only n possible values, the functional H becomes a function $H(\underset{\sim}{q},\underset{\sim}{p})$ of 2n variables $\underset{\sim}{q} = q_1, q_2, \ldots, q_n$ and $\underset{\sim}{p} = p_1, p_2, \ldots, p_n$. To show that H is equivalent to a symmetric function, let π be any permutation. By invariance, the minima of H and πH must coincide, where

$$\pi H(q,p) = H(q_{\pi(1)}, \ldots, q_{\pi(n)}, p_{\pi(1)}, \ldots, p_{\pi(n)}) \quad .$$

Therefore the minima of H and F coincide, where F is the mean of the πH for all permutations π, and H is equivalent to the symmetric function F. This completes the proof of Lemma II.

The subset independence property (Lemma I) and the symmetry of H in the discrete case (Lemma II) together enable us to prove that H is equivalent to functions that have the discrete sum form.

Theorem I: In the discrete case, let $H(\underset{\sim}{q},\underset{\sim}{p})$ satisfy uniqueness, invariance, and subset independence. Then H is equivalent to a function of the form

$$F(q,p) = \sum_j f(q_j,p_j) \tag{22}$$

for some function f of two variables.

Theorem I is proved in Appendix A. The proof rests primarily on the subset independence property (Lemma I).

We return to the continuous case by taking the limit of a sufficiently large number of sufficiently small subspaces $\underset{\sim}{S}_i$. The discrete sum form (22) then becomes

$$F(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\ f(q(\underset{\sim}{x}),p(\underset{\sim}{x})) \qquad (23)$$

C. Consequence of General Invariance in the Continuous Case

Although invariance was invoked for the special case of discrete permutations in deriving (22), the continuous sum form (23) does not satisfy the invariance axiom for arbitrary continuous transformations and arbitrary functions f. The invariance axiom restricts the possible forms of f as follows:

Theorem II: Let the functional H(q,p) satisfy uniqueness, invariance, and subset independence. Then H is equivalent to a functional of the form

$$F(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\ q(\underset{\sim}{x})h(q(\underset{\sim}{x})/p(\underset{\sim}{x})) \qquad (24)$$

for some function h of one variable.

Before proving Theorem II, we note that it illustrates the difficulty of dealing with an axiomatic characterization of the $\circ$ operator in comparison to an axiomatic characterization of H. If we knew that H itself must be transformation invariant, the deduction of (24) from (23) would be direct. But we know only that the minima of H must be transformation invariant. We suspect that the invariance axiom implies the existence of equivalent functionals that are themselves transformation invariant --- this is suggested by the proof that H can be assumed symmetric in the discrete case --- but we

have not been able to prove it. The following proof of Theorem II therefore reasons in terms of invariance at the minima of H.

Proof of Theorem II: From previous results we know that H may be assumed to have the form (23). Consider the case in which the new information I consists of a single equality constraint

$$\int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q^{\dagger}(\underset{\sim}{x})a(\underset{\sim}{x}) \;=\; 0. \tag{25}$$

Then, using standard techniques from the calculus of variations, it follows that the posterior $q = p \circ I$ satisfies

$$\lambda + \alpha a(\underset{\sim}{x}) \;+ g(q(\underset{\sim}{x}),p(\underset{\sim}{x})) \;=\; 0, \tag{26}$$

where λ and α are Lagrangian multipliers corresponding to the normalization constraint (2) and to (25) respectively, and where the function g is defined as

$$g(a,b) = \frac{\partial}{\partial a} f(a,b). \tag{27}$$

Now, let Γ be a coordinate transformation from $\underset{\sim}{x}$ to $\underset{\sim}{y}$ in the notation of Axiom II. Let $q' = \Gamma q$ for any $q \in \mathcal{Q}$. Then the transformed prior $p'(\underset{\sim}{y})$ and constraint function $a'(\underset{\sim}{y})$ are

$$p'(\underset{\sim}{y}) = \underset{\sim}{J}^{-1}\; p(\underset{\sim}{x}) \tag{28}$$

and

$$a'(\underset{\sim}{y}) = a(\underset{\sim}{x}), \tag{29}$$

where $\underset{\sim}{J}$ is the Jacobian of the transformation. The transformed constraints ΓI are

$$\int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q^{\dagger}(\underset{\sim}{x})a(\underset{\sim}{x}) \;=\; \int_{\underset{\sim}{D}} d\underset{\sim}{x}\,\underset{\sim}{J}\, q^{\dagger\prime}(\underset{\sim}{y})a'(\underset{\sim}{y}) \;=\; \int_{\underset{\sim}{D}'} d\underset{\sim}{y}\; q^{\dagger\prime}(\underset{\sim}{y})a'(\underset{\sim}{y}) = 0$$

and

$$\int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q^{\dagger}(\underset{\sim}{x}) \;=\; \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; \underset{\sim}{J} q^{\dagger\prime}(\underset{\sim}{y}) \;=\; \int_{\underset{\sim}{D}'} d\underset{\sim}{y}\; q^{\dagger\prime}(\underset{\sim}{y}) \;=\; 1.$$

The posterior $q' = p' \circ (\Gamma I)$, which is obtained by minimizing

$$H(q',p') = \int_{\underset{\sim}{D}'} d\underset{\sim}{y}\; f(q'(\underset{\sim}{y}),p'(\underset{\sim}{y})) \;,$$

satisfies

$$\lambda' \;+\; \alpha' a'(\underset{\sim}{y}) \;+ g(q'(\underset{\sim}{y}),p'(\underset{\sim}{y})) \;=\; 0, \qquad (30)$$

where λ' and α' are Lagrangian multipliers. Invariance (10) requires that the two posteriors be related by $q'(\underset{\sim}{y}) = \underset{\sim}{J}^{-1} q(\underset{\sim}{x})$, so (30) becomes

$$\lambda' \;+\; \alpha' a(\underset{\sim}{x}) \;+ g(q(\underset{\sim}{x})\underset{\sim}{J}^{-1}, p(\underset{\sim}{x})\underset{\sim}{J}^{-1}) \;=\; 0, \qquad (31)$$

where we have also used (28) and (29). Combining (26) and (31) yields

$$g(q(\underset{\sim}{x})\underset{\sim}{J}^{-1}, p(\underset{\sim}{x})\underset{\sim}{J}^{-1}) \;= g(q(\underset{\sim}{x}),p(\underset{\sim}{x})) + (\alpha - \alpha')a(\underset{\sim}{x}) + \lambda - \lambda' \;. \qquad (32)$$

Now, let $\underset{\sim}{S}_1, \ldots, \underset{\sim}{S}_n$ be disjoint subsets whose union is $\underset{\sim}{D}$ and let the prior p be constant within each $\underset{\sim}{S}_j$. If follows from Lemma II that q is also constant within each $\underset{\sim}{S}_j$, which in turn results in the right side of (32) being constant within each $\underset{\sim}{S}_j$. (The primed Lagrangian multipliers may depend on the transformation Γ, but they are constants.) On the left side, however, the Jacobian $\underset{\sim}{J}$ may take on arbitrary values since Γ is an arbitrary transformation. It follows that g can only depend on the ratio of its arguments, i.e., $g(a,b) = g(a/b)$. Eq. (27) then becomes

$$g(a/b) \;=\; \frac{\partial}{\partial a} f(a,b),$$

which has the general solution

$$f(a,b) = a\, h(a/b) + v(b), \qquad (33)$$

where h is some function of the ratio a/b and v is any function of b. Substitution of (33) into (24) yields

$$F(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q(\underset{\sim}{x})\; h(q(\underset{\sim}{x})/p(\underset{\sim}{x})) \;+\; \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; v(p(\underset{\sim}{x})) \quad .$$

Since the second term is a function only of the fixed prior, it cannot affect minimization of F and may be dropped. This completes the proof of Theorem II.

We note that, since $g(a,b) = g(a/b)$ holds, it follows from (32) that

$$(\alpha - \alpha')a(\underset{\sim}{x}) + \lambda - \lambda' = 0 \quad .$$

Since $a(\underset{\sim}{x})$ can be chosen as an arbitrary function, this shows that $\lambda = \lambda'$ and $\alpha = \alpha'$, i.e., the Lagrangian multipliers have values that are independent of the coordinate system.

D. Consequence of System Independence

Our results up to this point have not depended on Axiom III (system independence). We now show that system independence restricts the function h in (24) to a single equivalent form.

Theorem III. Let the functional H(q,p) satisfy uniqueness, invariance, subset independence, and system independence. Then H is equivalent to the functional

$$F(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q(\underset{\sim}{x}) \log(q(\underset{\sim}{x})/p(\underset{\sim}{x})) \quad , \qquad (34)$$

i.e., to cross-entropy.

Proof: Consider a system with states $\underset{\sim}{x}_1 \in \underset{\sim}{D}_1$, unknown density $q_1' \in Q_1$, prior density $p_1 \in Q_1$, and new information I_1 in the form of a single

equality constraint

$$\int_{\underset{\sim}{D}_1} d\underset{\sim}{x}_1 \; q_1(\underset{\sim}{x}_1)a(\underset{\sim}{x}_1) \;=\; 0 \quad . \tag{35}$$

From Theorem II, we may assume that H has the form (24). It follows that the posterior $q_1 = p_1 \circ I_1$ satisfies

$$\lambda_1 \;+\; \alpha_1 a(x_1) \;+\; u(r_1(x_1)) \;=\; 0 \;, \tag{36}$$

where λ_1 and α_1 are Lagrangian multipliers corresponding to the constraints (2) and (35), where $r_1(\underset{\sim}{x}_1) = q_1(\underset{\sim}{x}_1)/p_1(\underset{\sim}{x}_1)$, and where

$$u(r) \;=\; h(r) + r\,\frac{\partial}{\partial r} h(r) \quad . \tag{37}$$

Now consider another system with states $x_2 \in \underset{\sim}{D}_2$, unknown density $q_2^{\dagger} \in \mathcal{D}_2$, prior density $p_2 \in \mathcal{D}_2$, and new information I_2 in the form of a single equality constraint

$$\int_{\underset{\sim}{D}_2} d\underset{\sim}{x}_2 \; q_2(\underset{\sim}{x}_2)b(\underset{\sim}{x}_2) \;=\; 0 \quad . \tag{38}$$

The posterior $q_2 = p_2 \circ I_2$ satisfies

$$\lambda_2 \;+\; \beta_2 b(\underset{\sim}{x}_2) \;+\; u(r_2(\underset{\sim}{x}_2)) \;=\; 0 \;, \tag{39}$$

where λ_2 and β_2 are Lagrangian multipliers corresponding to the constraints (2) and (38), and where $r_2(\underset{\sim}{x}_2) = q_2(\underset{\sim}{x}_2)/p_2(\underset{\sim}{x}_2)$.

The two systems can also be described in terms of a joint probability density $q_{12}^{\dagger} \in \mathcal{D}_{12}$, a joint prior $p_{12} = p_1 p_2$, and new information I_{12} in the form of the three constraints

$$\int_{\underset{\sim}{D}_{12}} d\underset{\sim}{x}_1 d\underset{\sim}{x}_2 \; q_{12}(\underset{\sim}{x}_1, \underset{\sim}{x}_2) = 1 , \tag{40}$$

$$\int_{\underset{\sim}{D}_{12}} d\underset{\sim}{x}_1 d\underset{\sim}{x}_2 \; q_{12}(\underset{\sim}{x}_1, \underset{\sim}{x}_2) a(\underset{\sim}{x}_1) = 0 , \tag{41}$$

and

$$\int_{\underset{\sim}{D}_{12}} d\underset{\sim}{x}_1 d\underset{\sim}{x}_2 \; q_{12}(\underset{\sim}{x}_1, \underset{\sim}{x}_2) b(\underset{\sim}{x}_2) = 0 . \tag{42}$$

The posterior $q_{12} = p_{12} \circ I_{12}$ satisfies

$$\lambda_{12} + \alpha_{12} a(\underset{\sim}{x}_1) + \beta_{12} b(\underset{\sim}{x}_2) + u(r_{12}(\underset{\sim}{x}_1, \underset{\sim}{x}_2)) = 0 , \tag{43}$$

where the multipliers λ_{12}, α_{12}, and β_{12} correspond to (40)-(42), respectively, and where $r_{12} = q_{12}(\underset{\sim}{x}_1, \underset{\sim}{x}_2)/p_{12}(\underset{\sim}{x}_1, \underset{\sim}{x}_2)$.

Now, system independence (12) requires that $q_{12} = q_1 q_2$ hold, from which it follows that $r_{12} = r_1 r_2$ holds. Combining (36), (39) and (43) therefore yields

$$u(r_1 r_2) - u(r_1) - u(r_2) = (\alpha_1 - \alpha_{12})a + (\beta_1 - \beta_{12})b + \lambda_1 + \lambda_2 - \lambda_{12} .$$

Consider the case when $\underset{\sim}{D}_1$ and $\underset{\sim}{D}_2$ are both the real line. Then, differentiating this equation with respect to $\underset{\sim}{x}_1$ results in

$$u'(r_1 r_2) r_1' r_2 - u'(r_1) r_1' = (\alpha_1 - \alpha_{12}) a' ,$$

and differentiating this result with respect to $\underset{\sim}{x}_2$ yields

$$u''(r_1 r_2) r_1 r_2 + u'(r_1 r_2) = 0 . \tag{44}$$

By suitable choices for the priors and the constraints, $r_1 r_2$ can be made to take on any arbitrary positive value s. It follows from (44) that the function u satisfies the differential equation

$$\frac{du}{ds} + s\frac{d^2u}{ds^2} = 0,$$

which has the general solution

$$u(s) = A \log(s) + B, \tag{45}$$

for arbitrary constants A and B.

From (45) and (37), we obtain the following differential equation for the function h:

$$h(r) + r\frac{d}{dr}h(r) = A \log(r) + B . \tag{46}$$

Let us define $h_1(r) = r\, h(r)$. Then h_1 satisfies

$$\frac{dh_1}{dr} = A \log(r) + B .$$

The general solution for h_1 is $h_1(r) = A(r \log(r) - r) + Br + C$, so that the general solution of (46) is

$$h(r) = A\log(r) + C/r + B - A . \tag{47}$$

Substitution of (47) into (24) yields

$$F(q,p) = A\int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q(\underset{\sim}{x})\log(q(\underset{\sim}{x})/p(\underset{\sim}{x})) + (C + B - A) , \tag{48}$$

since the prior p satisfies the normalization constraint (2). Since the constants A, B, and C cannot affect the minimization of (48), provided $A > 0$, this completes the proof of Theorem III.

E. Cross-Entropy Satisfies the Axioms

So far, we have shown that, if H(q,p) satisfies the axioms, then H is

equivalent to cross-entropy (24). This still leaves open the possibility that no functional H satisfies the axioms for arbitrary constraints. By showing that cross-entropy satisfies the axioms for arbitrary constraints, we complete the proof of our main result:

Theorem IV: The cross-entropy

$$H(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q(\underset{\sim}{x})\log(q(\underset{\sim}{x})/p(\underset{\sim}{x})) \tag{49}$$

satisfies uniqueness, invariance, system independence, and subset independence. Every other functional that satisfies the axioms is equivalent to cross-entropy.

Proof: We need only show that cross-entropy satisfies the axioms.

Uniqueness. Let $\mathcal{J}$ be any closed, convex subset $\mathcal{J} \subseteq \mathcal{Q}$, and let densities $q,r \in \mathcal{J}$ have the same cross entropy $H(q,p) = H(r,p)$ for some prior $p \in \mathcal{Q}$. We define $g(u) = u\log(u)$, with $g(0) = 0$, so that H can be written as

$$H(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; p(\underset{\sim}{x})g(q(\underset{\sim}{x})/p(\underset{\sim}{x})) \quad .$$

Now, since $g''(u) = 1/u > 0$, g is strictly convex. It follows that

$$\alpha g(u) + (1-\alpha)g(v) > g(\alpha u + (1-\alpha)v) \ ,$$

for $0 < \alpha < 1$ and $u \neq v$. We can therefore write

$$\begin{aligned} H(q,p) &= H(r,p) \\ &= \int_{\underset{\sim}{D}} d\underset{\sim}{x}\left[\alpha p(x) g\left(\frac{q(\underset{\sim}{x})}{p(\underset{\sim}{x})}\right) + (1-\alpha)p(\underset{\sim}{x})g\left(\frac{r(\underset{\sim}{x})}{p(\underset{\sim}{x})}\right)\right] \\ &\geqslant \int_{\underset{\sim}{D}} d\underset{\sim}{x}\; p(\underset{\sim}{x})g\left(\frac{\alpha q(\underset{\sim}{x}) + (1-\alpha)r(\underset{\sim}{x})}{p(\underset{\sim}{x})}\right) \end{aligned} \tag{50}$$

The inequality is strict unless $q = r$. (We write $q = r$ when $q(\underset{\sim}{x}) = r(\underset{\sim}{x})$ for all $\underset{\sim}{x}$ except at most a set of measure zero, since in this case q and r define the same probability distribution and we should not distinguish between them.) Eq. (50) shows that, if $q \neq r$ and $H(q,p) = H(r,p)$ both hold, there exists a density $s = \alpha q + (1-\alpha)r$ that satisfies $s \in \mathcal{J}$ (since $\mathcal{J}$ is convex) and has smaller cross entropy $H(s,p) < H(q,p)$. Therefore, there cannot be two distinct densities $q,r \in \mathcal{J}$ having the minimum cross-entropy in $\mathcal{J}$. This proves that cross-entropy satisfies Axiom I.

Invariance. Let Γ be a coordinate transformation from $\underset{\sim}{x}$ to $\underset{\sim}{y}$ in the notation of Axiom II. Let $q' = \Gamma q$ for any $q \in \mathcal{D}$. Then

$$H(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\, q(\underset{\sim}{x}) \log(q(\underset{\sim}{x})/p(\underset{\sim}{x})) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\, J\, q'(\underset{\sim}{y}) \log(q'(\underset{\sim}{y})/p'(\underset{\sim}{y}))$$

$$= \int_{\underset{\sim}{D}'} d\underset{\sim}{y}\, q'(\underset{\sim}{y}) \log(q'(\underset{\sim}{y})/p'(\underset{\sim}{y}))$$

shows that cross-entropy is transformation invariant. The minimum in $\Gamma\mathcal{J}$ therefore corresponds to the minimum in $\mathcal{J}$, which proves that cross-entropy satisfies Axiom II.

System Independence. We use the notation in Axiom III. Consider densities $q_1, p_1 \in \mathcal{D}_1$ and $q_2, p_2 \in \mathcal{D}_2$. Let the density $q \in \mathcal{D}_{12}$ satisfy $q \neq q_1 q_2$,

$$\int_{\underset{\sim}{D}_1} d\underset{\sim}{x}_1\, q(\underset{\sim}{x}_1, \underset{\sim}{x}_2) = q_2 \quad ,$$

and

$$\int_{\underset{\sim}{D}_2} d\underset{\sim}{x}_2\, q(\underset{\sim}{x}_1, \underset{\sim}{x}_2) = q_1 \quad ,$$

i.e., q and $q_1 q_2$ are different densities with the same marginal densities. We compute the cross-entropy difference between q and $q_1 q_2$ for

the same prior p_1p_2 as follows:

$$H(q,p_1p_2) - H(q_1q_2,p_1p_2) =$$

$$\iint_{\underset{\sim}{D}_1\underset{\sim}{D}_2} d\underset{\sim}{x}_1 d\underset{\sim}{x}_2\; q(\underset{\sim}{x}_1,\underset{\sim}{x}_2)\left[\log\left(\frac{q(\underset{\sim}{x}_1,\underset{\sim}{x}_2)}{p_1(\underset{\sim}{x}_1)p_2(\underset{\sim}{x}_2)}\right) - \log\left(\frac{q_1(\underset{\sim}{x}_1)}{p_1(\underset{\sim}{x}_1)}\right) - \log\left(\frac{q_2(\underset{\sim}{x}_2)}{p_2(\underset{\sim}{x}_2)}\right)\right]$$

$$= \iint d\underset{\sim}{x}_1 d\underset{\sim}{x}_2\; q(\underset{\sim}{x}_1,\underset{\sim}{x}_2)\log\left(\frac{q(\underset{\sim}{x}_1,\underset{\sim}{x}_2)}{q_1(\underset{\sim}{x}_1)q_2(\underset{\sim}{x}_2)}\right)$$

$$= H(q,q_1q_2) \quad .$$

Now, cross-entropy has the property that $H(q,p) \geq 0$ with $H(q,p) = 0$ only if $q = p$ (for example, see [31, p. 14]). It follows that

$$H(q,p_1p_2) > H(q_1q_2,p_1p_2) \tag{51}$$

holds, since $q \neq q_1q_2$ by assumption. Eq. (51) states that, of all the densities $q \in \mathcal{D}_{12}$ with given marginal densities q_1 and q_2, the one with least cross-entropy is q_1q_2. Since I_1 and I_2 restrict only the marginal densities of q in $q = (p_1p_2)\circ(I_1 \wedge I_2)$ --- see Axiom III and the text preceding it --- the density q with the least cross-entropy in the constraint set is of the product form q_1q_2. But the cross-entropy of a density of this form is given by

$$H(q_1q_2,p_1p_2) = \iint_{\underset{\sim}{D}_1\underset{\sim}{D}_2} d\underset{\sim}{x}_1 d\underset{\sim}{x}_2\; q_1(\underset{\sim}{x}_1)q_2(\underset{\sim}{x}_2)\log\left(\frac{q_1(\underset{\sim}{x}_1)\,q_2(\underset{\sim}{x}_2)}{p_1(\underset{\sim}{x}_1)\,p_2(\underset{\sim}{x}_2)}\right)$$

$$= \int_{\underset{\sim}{D}_1} d\underset{\sim}{x}_1\; q_1(\underset{\sim}{x}_1)\log\left(\frac{q_1(\underset{\sim}{x}_1)}{p_1(\underset{\sim}{x}_1)}\right) + \int_{\underset{\sim}{D}_2} d\underset{\sim}{x}_2\; q_2(\underset{\sim}{x}_2)\log\left(\frac{q_2(\underset{\sim}{x}_2)}{p_2(\underset{\sim}{x}_2)}\right)$$

$$= H(q_1,p_1) + H(q_2,p_2) \quad , \tag{52}$$

and so assumes its minimum when the two terms on the right assume their individual minima --- the first subject to I_1 and the second to I_2. Thus, we have $q = (p_1 p_2) \circ (I_1 \wedge I_2) = q_1 q_2 = (p_1 \circ I_1)(p_2 \circ I_2)$, and we have proven that cross-entropy satisfies Axiom III.

Subset Independence. We use the notation in Axiom IV. We also define $q = p \circ (I \wedge M)$, $q_i = q^* \underset{\sim}{S}_i$, and $p_i = p^* \underset{\sim}{S}_i$. (Eq. (15) then becomes $q_i = p_i \circ I_i$.) The cross-entropy of q with respect to p may be written

$$H(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\, q(\underset{\sim}{x}) \log(q(\underset{\sim}{x})/p(\underset{\sim}{x}))$$

$$= \sum_i \int_{\underset{\sim}{S}_i} d\underset{\sim}{x}\, m_i q_i(\underset{\sim}{x}) \log\left(\frac{m_i q_i(\underset{\sim}{x})}{s_i p_i(\underset{\sim}{x})}\right)$$

$$= \sum_i \int_{\underset{\sim}{S}_i} d\underset{\sim}{x}\, q_i(\underset{\sim}{x}) \log\left(\frac{q_i(\underset{\sim}{x})}{p_i(\underset{\sim}{x})}\right) + \sum_i m_i \log\left(\frac{m_i}{s_i}\right)$$

$$= \sum_i H(q_i,p_i) + \sum_i m_i \log\left(\frac{m_i}{s_i}\right), \qquad (53)$$

where the s_i are the prior probabilities of being in each subset

$$s_i = \int_{\underset{\sim}{S}_i} d\underset{\sim}{x}\, p(\underset{\sim}{x}) .$$

The second sum on the right of (53) is a constant and has no effect on minimization. Minimizing the left side of (53) subject to $(I \wedge M)$ is equivalent to minimizing each term of $\sum_i H(q_i,p_i)$ individually subject to I_i. This proves that cross-entropy satisfies subset independence and completes the proof of Theorem IV. For a discussion of the mathematics of minimizing cross-entropy (49) subject to constraints, see Appendix B.

V. THE DISCRETE CASE

A. Principle of Minimum Cross-Entropy for Discrete Systems

Theorem IV states that, if one wishes to select a posterior $q = p \circ I$ in a manner that satisfies Axioms I-IV, the unique result can be obtained by minimizing the cross-entropy (24). Although the equivalent result for the discrete case can be obtained in the usual informal way by replacing integrals with sums and densities with distributions, it can also be obtained formally as follows.

Suppose a system has a finite set of n states with probabilities $\underset{\sim}{q}^\dagger = q_1^\dagger,..,q_n^\dagger$. Let $\underset{\sim}{p} = p_1,..,p_n$ be a prior estimate of $\underset{\sim}{q}^\dagger$ and let new information be provided in the form

$$\sum_i q_i^\dagger a_{ki} = 0 \tag{54}$$

or

$$\sum_i q_i^\dagger c_{ki} \geqslant 0 \ , \tag{55}$$

for known sets of numbers a_{ki} and c_{ki}. Then it is clear that there exist problems with continuous states and densities for which the foregoing finite problem is the discrete case as defined in Lemma II. It follows from Lemma II and Theorem IV that the cross entropy functional becomes a function of 2n variables and that the posterior $\underset{\sim}{q} = q_1,\ldots,q_n$ can be obtained by minimizing the function

$$H(\underset{\sim}{q},\underset{\sim}{p}) = \sum_i q_i \log(q_i/p_i) \ , \tag{56}$$

subject to the constraints (54)-(55).

B. The Maximum Entropy Principle

Using transformation group arguments, Jaynes [25] has shown that a uniform prior $p_i = n^{-1}$ is appropriate when we know only that each of the n system states is possible (as distinct from "complete ignorance" when we don't even know this much). It follows that, given only a finite set of possible states and new information in the form of discrete constraints (54)-(55), the posterior is obtained by minimizing the function

$$H(q) = \sum_i q_i \log(q_i) - \log(n).$$

This is equivalent to maximizing the entropy $-\sum_i q_i \log(q_i)$. We conclude that the principle of maximum entropy is a special case of our general results for cross-entropy minimization.

It is also possible to obtain the maximum entropy principle formally and directly. We show how in the following, although we omit some of the formal details. The first step is to rewrite the axioms so that they refer to the discrete case in which no prior is available. In this case, given new information I in the form of constraints (54)-(55), the unary operator $\circ$ selects a posterior distribution $q = (\circ I)$ from all distributions that satisfy the constraints. The operator is realized by minimizing some function $H(\underset{\sim}{q})$. The axioms become (see Section III):

I (uniqueness): The posterior $\underset{\sim}{q} = (\circ I)$ is unique.

II (permutation invariance): $\circ(\Gamma I) = \Gamma(\circ I)$ for any permutation Γ. (57)

III (system independence): $\circ(I_1 \wedge I_2) = (\circ I_1)(\circ I_2)$.

IV (subset independence): $(\circ(I \wedge M))*\underset{\sim}{S}_i = (\circ I_i)$.

Theorem I goes through in a straightforward way with the prior deleted. This shows that, if $H(\underset{\sim}{q})$ satisfies uniqueness, permutation invariance, and subset independence, it is equivalent to a function of the form

$$H(\underset{\sim}{q}) = \sum_i f(q_i) \ . \tag{58}$$

We now assume the form (58) and apply system independence in a manner analogous to the proof of Theorem III. Consider a system with n states, unknown distribution $\underset{\sim}{q}^\dagger$, and new information I_1 in the form of a single equality constraint

$$\sum_{i=1}^{n} q_i^\dagger a_i = 0 \ . \tag{59}$$

The posterior $\underset{\sim}{q} = (\bullet I_1)$ satisfies

$$u(q_i) + \alpha_1 a_i + \lambda_1 = 0 \tag{60}$$

$(i = 1,\ldots,n)$, where the function u is defined as

$$u(x) = \frac{d}{dx} f(x) \ , \tag{61}$$

and where α_1 and λ_1 are Lagrangian multipliers corresponding to (59) and the usual normalization constraint. Now consider a second system, this one with m states, an unknown distribution $\underset{\sim}{r}^\dagger$, and new information I_2 in the form of the constraint

$$\sum_{k=1}^{m} r_k^\dagger b_k = 0 \ . \tag{62}$$

The posterior $\underset{\sim}{r} = (\bullet I_2)$ satisfies

$$u(r_k) + \beta_2 b_k + \lambda_2 = 0 \tag{63}$$

$(k = 1,\ldots,m)$, where β_2 and λ_2 are Lagrangian multipliers corresponding to (62) and the normalization constraint. Since the two systems can be described in terms of a joint distribution, and since a joint posterior can be selected in accordance with both I_1 and I_2, the following equation also holds:

$$u(q_i r_k) + \alpha_{12} a_i + \beta_{12} b_k + \lambda_{12} = 0 \tag{64}$$

$(i = 1,\ldots,n,\ k = 1,\ldots,m)$. In (64) we have already applied the system independence axiom and written the joint posterior as the product of the individual posteriors. Combining (60), (63), and (64), yields

$$u(q_i r_k) = u(q_i) + u(r_k) + (\alpha_1 - \alpha_{12})a_i + (\beta_2 - \beta_{12})b_k$$

This leads to

$$\begin{aligned} u(q_i r_k) - u(q_i r_v) &= u(q_u r_k) - u(q_u r_v) \\ &= G(r_k, r_v), \end{aligned} \tag{65}$$

for some function G. Since the right side of (65) does not depend on q_i, we pick an arbitrary value for q_i on the left side. This shows that G satisfies

$$G(x,y) = s(x) - s(y) \tag{66}$$

for some function s. (We note that G satisfies Sincov's functional equation $G(x,y) = G(x,z) + G(z,y)$, which has the general solution (66) [56, p. 223].) Some manipulation of (65) and (66) yields

$$u(xy) - s(x) - s(y) = u(wz) - s(w) - s(z) \tag{67}$$

Both sides of (67) are independent of each other and must therefore be equal to some constant, so that u satisfies $u(xy) = g(x) + g(y)$, for some function

g. Using standard techniques of functional equations [56, pp. 34,302], we obtain the general solution for u(x), namely u(x) = A log(x) + B, where A and B are constants. Combining this solution with (61) and integrating yields the solution for f in (58), f(x) = A x log(x) + Bx - A, which in turn yields

$$H(q) = A \sum_i q_i \log(q_i) - nA + B \quad . \tag{68}$$

This function has a unique minimum provided that A is positive.

Minimizing the function H in (68) is equivalent to maximizing the entropy $-\sum_i q_i \log(q_i)$. This proves that, if one wishes to select a discrete posterior distribution q = (•I) in a manner that satisfies the axioms (57), the unique result can be obtained by maximizing entropy.

VI. INFERENCE AXIOMS VS. INFORMATION MEASURE AXIOMS

Our approach has been to axiomatize desired properties of inference methods rather than to axiomatize desired properties of information measures. Yet it might seem that the axioms in Section III are no more than a thinly-disguised characterization of cross-entropy. In this view, Axioms I and II might correspond to axioms requiring that H have unique minima and be transformation invariant, and Axioms III and IV might correspond to axioms requiring that H be "additive" [34] and satisfy something like the "branching property" [57]. These correspondences are meaningful and not surprising --- after all, inference methods should relate to information measures --- but it is important to realize that there are significant differences as well.

Consider Axiom III (system independence) and the following axiom, which can be used [34] in characterizing the directed divergences $\int d\underset{\sim}{x}\ q(\underset{\sim}{x})\log(q(\underset{\sim}{x})/p(\underset{\sim}{x}))$ (cross-entropy) and $\int d\underset{\sim}{x}\ p(\underset{\sim}{x})\log(p(\underset{\sim}{x})/q(\underset{\sim}{x}))$:

Additivity. $H(q_1q_2,p_1p_2) = H(q_1,p_1) + H(q_2,p_2)$ (69)

for all $q_1,p_1 \in \mathcal{Q}_1$ and $q_2,p_2 \in \mathcal{Q}_2$.

In Section IV we showed that, if H has the form

$$H(q,p) = \int_{\underset{\sim}{D}} d\underset{\sim}{x}\ q(\underset{\sim}{x})h(q(\underset{\sim}{x})/p(\underset{\sim}{x})) \tag{70}$$

and is required to satisfy system independence, then it follows that H is equivalent to cross-entropy (Theorem III). When we proved, as part of Theorem IV, that cross-entropy itself satisfies system independence, we exploited the fact that cross-entropy satisfies additivity (69) --- see (52). At first glance, it might seem that any functional that satisfies additivity also satisfies system independence. But Johnson [34] proved that the information

measures $H(q,p)$ of the form (70) that satisfy additivity (69) are those of the form

$$H(q,p) = A\int_{\underset{\sim}{D}} d\underset{\sim}{x}\; q(\underset{\sim}{x})\log(q(\underset{\sim}{x})/p(\underset{\sim}{x})) + B\int_{\underset{\sim}{D}} d\underset{\sim}{x}\; p(\underset{\sim}{x})\log(p(\underset{\sim}{x})/q(\underset{\sim}{x})) , \quad (71)$$

for some constants $A,B \geq 0$, not both zero. That is, (70) and additivity (69) of H yields the linear combination of both directed divergences, whereas (70) and system independence of the operator $\circ$ yields only one of the directed divergences, cross-entropy. The key to the difference is the property expressed by (51) --- for all densities $q \in \mathcal{D}_{12}$ with given marginal densities q_1 and q_2, $H(q,p_1p_2)$ has its minimum at $q = q_1q_2$. This property is necessary if H is to satisfy system independence; it is satisfied by the first term in (71) but not by the second, even though the second term satisfies additivity.

VII. SUMMARY

We have proved that, in a well-defined sense, Jaynes's principle of maximum entropy and Kullback's principle of minimum cross-entropy (minimum directed divergence) provide correct, general methods of inductive inference when given new information in the form of expected values. In contrast to previous justifications, our approach has been to axiomatize desired properties of inference methods rather than to axiomatize desired properties of information measures. We defined four axioms, all of them based on the principle that, if a problem can be solved in more than one way, the results should be consistent. We proved that, when given a prior continuous density or discrete distribution and new information in the form of expected values, there is only one posterior density that can be chosen in a manner that satisfies the consistency axioms. This unique posterior can be obtained by minimizing cross-entropy subject to the constraints of the new information. In the discrete case when no prior is available, we proved that there is still only one posterior that can be chosen in a manner that satisfies the consistency axioms, and that this unique posterior can be obtained by maximizing entropy. This result for the principle of maximum entropy was obtained both from the general result for minimum cross-entropy, in the special case of uniform priors, and directly as a consequence of prior-free versions of the axioms.

When Jaynes first proposed the maximum entropy principle more than 20 years ago, he did not ignore such questions as "Why maximize entropy, why not some other function?". We have confirmed his conjecture [1, p. 623] that "deductions made from any other information measure, if carried far enough, will eventually lead to contradictions."

APPENDIX A

Proof of Theorem I

The proof has three steps. First, we discuss some general properties of H and we show that the partial derivatives of H have the form

$$\frac{\partial H}{\partial q_j} = z(\underset{\sim}{q},\underset{\sim}{p}) + g(q_j,p_j)s(\underset{\sim}{q},\underset{\sim}{p}) \qquad (A.1)$$

for some functions z,g, and s. This follows from invariance and subset independence. We then investigate the behavior of g and s and show that (A.1) results in H being functionally dependent on $F(\underset{\sim}{q},\underset{\sim}{p}) = \sum_j f(q_j,p_j)$, where f is a solution of $g = \partial f(q,p)/\partial q$. Finally, we show that the functional dependence is monotonic so that H and F are equivalent.

In realizing the operator $\circ$, the only relevant values of $H(\underset{\sim}{q},\underset{\sim}{p})$ are at points $\underset{\sim}{q}$ that satisfy the normalization constraint

$$\sum_{j=1}^{n} q_j = 1 \quad . \qquad (A.2)$$

This is just the discrete form of (1), which is required by $q \in Q$ (see Section II). We shall refer to the hyperplane consisting of points $\underset{\sim}{q}$ that satisfy (A.2) as the normalization subspace. In selecting posteriors by minimizing H, we are further restricted to the positive region of points $\underset{\sim}{q}$ that satisfy $q_i \geqslant 0$ for $i = 1,\ldots,n$. This restriction is also required by $q \in Q$. On the normalization subspace (A.2), $H(\underset{\sim}{q},\underset{\sim}{p})$ is a function of only n-1 independent variables q_i (the prior $\underset{\sim}{p}$ is assumed fixed). For convenience, however, we consider H to be extended off the normalization subspace to a well-behaved function of n independent variables that is symmetric under identical permuations of $\underset{\sim}{q}$ and $\underset{\sim}{p}$ (see Lemma II). This enables us to express the

gradient $\underset{\sim}{\nabla} H$ as

$$H = \sum_{i=1}^{n} \frac{\partial H}{\partial q_i} \hat{\underset{\sim}{e}}_i \quad ,$$

where $\hat{\underset{\sim}{e}}_1, \ldots, \hat{\underset{\sim}{e}}_n$ is a standard, orthonormal basis. The operator $\circ$ can be realized by minimizing the extended H in the positive region provided that (A.2) is always imposed as a constraint. In the continuous case, we have assumed that the functional H(q,p) is well-behaved. We take this to mean, in particular, that the function $H(\underset{\sim}{q},\underset{\sim}{p})$ is continuously differentiable in the interior of the positive region of the normalization subspace and that the projection of $\underset{\sim}{\nabla} H$ into the normalization subspace is zero only at minima of H.

Now, let N be the set of integers $\{1,\ldots n\}$, let $M \subset N$ be a set of m integers from N, and let M-N be the set that remains after deleting M from N. Let $\underset{\sim}{q}_M$ comprise the subset of components q_i with $i \in M$ and let $\underset{\sim}{q}_{N-M}$ comprise the rest. We shall refer to points $\underset{\sim}{q}_M$ as points in the M-subspace. In the derivation that follows, we assume that $n \geqslant 6$ and $m \geqslant 4$ both hold. Suppose new information comprises a set of equality constraints (20)

$$\sum_{j=1}^{n} q_j^{\dagger} a_{kj} = 0 \qquad \text{(A.3)}$$

that satisfy $a_{kj} = 0$ either for all $j \in M$ or for all $j \in N-M$, and suppose the set includes the constraint

$$\sum_{j \in M} q_j^{\dagger} = r. \qquad \text{(A.4)}$$

If a particular constraint satisfies $a_{kj} = 0$ for $j \in M$, then it can be written as a constraint

$$\sum_{j \in M} a_{kj} q_j^{\dagger} = \sum_{j \in M} a_{kj} (q_j^{\dagger}/r) = 0$$

on the conditional distribution ($\underset{\sim}{q}_M/r$), i.e., as a constraint on the conditional distribution given $j \in M$. Similarly, constraints that satisfy $a_{kj} = 0$ for $j \in N-M$ can be written as constraints on the conditional distribution $\underset{\sim}{q}_{N-M}/(1-r)$. Therefore, the system decomposes into two subsets (M and N-M) with new information that satisfies the assumptions of Axiom IV (subset independence). It follows from Lemma I that, when $H(\underset{\sim}{q},\underset{\sim}{p})$ is minimized over the constraint set, the resulting $\underset{\sim}{q}_M$ are independent of the $\underset{\sim}{q}_{N-M}$, of the $\underset{\sim}{p}_{N-M}$, and of n.

Now, although the M-subspace is m-dimensional, the constraint (A.4) requires that the solution $\underset{\sim}{q}_M$ be found on the m-1 dimensional hyperplane defined by (A.4). Therefore, finding this solution depends, not on the projection of $\underset{\sim}{\nabla} H$ into the M-subspace,

$$(\underset{\sim}{\nabla} H)_M = \sum_{j \in M} \frac{\partial H}{\partial q_j} \hat{\underset{\sim}{e}}_j \quad ,$$

but on its projection onto the (m-1) dimensional hyperplane defined by (A.4). This projection is given by

$$\underset{\sim}{B}_M = (\underset{\sim}{\nabla} H)_M - (\hat{\underset{\sim}{n}} \cdot (\underset{\sim}{\nabla} H)_M)\hat{\underset{\sim}{n}} \ ,$$

where $\hat{\underset{\sim}{n}}$ is a unit vector normal to the hyperplane. Since

$$\hat{\underset{\sim}{n}} = \frac{1}{\sqrt{m}} \sum_{j \in M} \hat{\underset{\sim}{e}}_j$$

will serve, $\underset{\sim}{B}_M$ has components

$$B_{Mi} = \frac{\partial H}{\partial q_i} - \frac{1}{m} \sum_{j \in M} \frac{\partial H}{\partial q_j} \qquad \text{(A.5)}$$

for $i \in M$. Now, since H is symmetric (Lemma II),

$$\frac{\partial H}{\partial q_i} = h(q_i, \underset{\sim}{q}_{N-i}, p_i, \underset{\sim}{p}_{N-i}) \equiv h_i$$

holds for some function h, where $\underset{\sim}{q}_{N-i}$ is any permutation of $q_1, \ldots, q_n$ with q_i deleted and $\underset{\sim}{p}_{N-i}$ is the same permuation of $p_1, \ldots, p_n$ with p_i deleted. Hence, (A.5) becomes

$$B_{Mi} = h_i - \frac{1}{m} \sum_{j \in M} h_j$$

$$= B(q_i, \underset{\sim}{q}_{N-i}, p_i, \underset{\sim}{p}_{N-i}) \quad ,$$

for some function B.

To find the solution for $\underset{\sim}{q}_M$, one moves on the constraint hyperplane opposite the direction of maximum change in H --- i.e., opposite the direction of $\underset{\sim}{B}_M$ --- until no further movement is possible within the constraint set (A.3). Since the solution cannot depend on $\underset{\sim}{q}_{N-M}$ or $\underset{\sim}{p}_{N-M}$, neither can the direction of $\underset{\sim}{B}_M$. This direction is also independent of n, since the subspace solution $\underset{\sim}{q}_M$ is independent of n (Lemma I). If $\underset{\sim}{U}_M$ is a unit vector in the direction of $\underset{\sim}{B}_M$, with components U_{Mi}, it follows that

$$U_{Mi} \equiv \frac{B_{Mi}}{|\underset{\sim}{B}_M|} = U(q_i, \underset{\sim}{q}_{M-i}, p_i, \underset{\sim}{p}_{M-i}) \quad , \tag{A.6}$$

for some function U, where $\underset{\sim}{q}_{M-i}$ is any permutation of $\underset{\sim}{q}_M$ with q_i deleted, etc. The function U is well defined everywhere on the constraint hyperplane except at a point at which H is minimized subject only to (A.4). Such a point is characterized equivalently by $\underset{\sim}{B}_M = 0$ and by $h_i = h_j$ for all $i,j \in M$. By uniqueness, there cannot be more than one such point. For if

there were more than one such point, H would would reach its minimum value at more than one point or would have local minima in addition to an absolute minimum. In either case, one could define convex constraint sets in which the minimum of H would occur at more than one point, thereby violating uniqueness.

The point at which (A.6) is ill-defined is also characterized by the equality of the ratios $(q_i/p_i)=(q_j/p_j)$ for all $i,j \in M$. To see this, we apply the subset independence axiom. Minimizing H subject only to (A.4) means that (15) applies without the additional information I. If we define b as

$$b = \sum_{j \in M} p_j ,$$

then (15) becomes $(q_j/r) = (p_j/b)$ so that q_j/p_j is a constant independent of j for $j \in M$. In the case of n = m, the constraint hyperplane becomes the entire positive region of the normalization subspace --- (A.4) becomes equivalent to (A.2) and r = b = 1 holds. This shows that there is only one point at which all of the h_i are equal, namely the point $\underset{\sim}{q} = \underset{\sim}{p}$. Similarly, by taking m = 2 and $M = \{i,j\}$, one can show that the condition $h_i = h_j$ is equivalent to the condition $(q_i/p_i)=(q_j/p_j)$.

From (A.6) we obtain

$$\frac{B_{Mi} - B_{Mj}}{B_{Mk} - B_{Mj}} = \frac{U_{Mi} - U_{Mj}}{U_{Mk} - U_{Mj}} \qquad (A.7)$$

for $i,j,k \in M$. But

$$\frac{B_{Mi} - B_{Mj}}{B_{Mk} - B_{Mj}} = \frac{h_i - h_j}{h_k - h_j} \qquad (A.8)$$

follows from (A.5). Since the right-hand side of (A.8) cannot depend on the

definition of M, neither can the right-hand side of (A.7). It follows that

$$\frac{h_i - h_j}{h_k - h_j} = W(q_i,q_j,q_k,p_i,p_j,p_k) \equiv W_{ijk} \qquad \text{(A.8)}$$

holds for some function W of six variables. Now, by this construction, W is well defined when $q_i+q_j+q_k < 1$ and $h_k \neq h_j$; the latter condition is equivalent to $(q_k/p_k)\neq(q_j/p_j)$. However,

$$\frac{h_i - h_j}{h_k - h_j} = \frac{W_{iju}}{W_{kju}} = W_{ijk}$$

holds, and further manipulation yields

$$\frac{W_{iru} - W_{jru}}{W_{kru} - W_{jru}} = W_{ijk} \ . \qquad \text{(A.9)}$$

Since (A.9) is independent of q_r, q_u, p_r, and p_u, we may take arbitrary values of these variables, and use (A.9) to extend the definition of W. By the discussion following (A.8), the numerator and denominator on the left of (A.9) are defined as long as $(q_r/p_r)\neq(q_u/p_u)$ holds and then the fraction is well defined whenever

$(q_k/p_k) \neq (q_j/p_j)$,

$0 \leq q_i < 1 - q_u - q_r$,

$0 \leq q_j < 1 - q_u - q_r$,

and

$0 \leq q_k < 1 - q_u - q_r$

all hold. But we can make $1 - q_u - q_r$ arbitrarily close to 1 so that we may extend the domain of W_{ijk} to include all arguments such that q_i, q_j,

and q_k are between 0 and 1 and $(q_k/p_k) \neq (q_j/p_j)$ holds. Moreover, (A.9) continues to hold on this extended domain.

Now we may write $g(q_i,p_i) \equiv g_i$ for W_{iru} with some particular, fixed values of q_r, q_u, and obtain

$$\frac{h_i - h_j}{h_k - h_j} = \frac{g(q_i,p_i) - g(q_j,p_j)}{g(q_k,p_k) - g(q_j,p_j)} \tag{A.10}$$

for some function g of two variables. It follows that $h_i = \partial H/\partial q_i$ has the form

$$h_i = z(\underset{\sim}{q},\underset{\sim}{p}) + g(q_i,p_i)s(\underset{\sim}{q},\underset{\sim}{p}) \tag{A.11}$$

for some functions z, s, and g. This implies that

$$h_i - h_j = (g_i - g_j)s \tag{A.12}$$

holds, so that the function s is given by

$$s = \frac{h_i - h_j}{g_i - g_j}.$$

For a particular point $\underset{\sim}{q}$, the right-hand side may be ill-defined for certain values of i and j. Since s is independent of i and j, however, s is well-defined unless $g_i = g_j$ for all i,j. But, from the construction of g, the condition $g_i = g_j$ is equivalent to $h_i = h_j$ and therefore to $(q_i/p_i)=(q_j/p_j)$. It follows that s is well-defined everywhere in the positive region of the normalization subspace except perhaps at the single point where $\underset{\sim}{q} = \underset{\sim}{p}$ holds. The function z is likewise well-defined except perhaps at this point.

Furthermore, s and g are continuous except perhaps at $\underset{\sim}{q} = \underset{\sim}{p}$: Since H is continuously differentiable, the derivatives h_i are continuous and finite everywhere in the positive region of the normalization subspace (except possibly on the boundary --- at points that satisfy $q_i = 0$ for some i). It follows that each of the functions $(\underset{\sim}{\nabla} H)_M$, $\underset{\sim}{B}_M$, U, W, s, and g is continuous except perhaps at certain "obvious" points where it is ill-defined because of a vanishing denominator in the construction.

Now, let t parameterize some curve $\underset{\sim}{q}(t)$ in the positive region of the normalization subspace. It follows from (A.11) that

$$\frac{d}{dt} H(\underset{\sim}{q}(t),\underset{\sim}{p}) = \sum_{i=1}^{n} \dot{q}_i \frac{\partial H}{\partial q_i}$$

$$= s \sum_i \dot{q}_i g_i + z \sum_i \dot{q}_i$$

holds, where $\dot{q}_i = dq_i/dt$. But the normalization constraint (A.2) implies that $\sum_i \dot{q}_i = 0$, so we have

$$\frac{d}{dt} H(\underset{\sim}{q}(t),\underset{\sim}{p}) = s(\underset{\sim}{q}(t),\underset{\sim}{p}) \frac{d}{dt} F(\underset{\sim}{q}(t),\underset{\sim}{p}) , \tag{A.13}$$

where

$$F(\underset{\sim}{q},\underset{\sim}{p}) = \sum_{i=1}^{n} f(q_i,p_i) \tag{A.14}$$

for some function f related to g by $g(q_i,p_i) = \partial f(q_i,p_i)/\partial q_i$. Suppose the curve $\underset{\sim}{q}(t)$ lies in a level surface of H. Then dH/dt = 0 and (A.13) shows that F will also be constant on any such curve, unless perhaps s is zero. However, (A.12) shows that, in the interior of the normalization subspace, s in not zero unless $h_i = h_j$ for all i,j, which can be true only at the point $\underset{\sim}{q} = \underset{\sim}{p}$. It therefore follows that F is constant on connected

components of level surfaces of H and that F and H are functionally dependent --- locally, F can be written as a function of H, with

$$\frac{dF}{dH} = \frac{1}{s(q,p)} \tag{A.15}$$

Next, we show that the functional dependence is monotonic. If it were not monotonic, then $dF/dH = s^{-1}(\underset{\sim}{q},\underset{\wedge}{p})$ would change sign at a point $\underset{\sim}{q}$ and, therefore, in some neighborhood of $\underset{\wedge}{q}$ along a level surface of H. But we have already shown that s is continuous and non-zero everywhere in the interior of the normalization subspace except perhaps at a single point (the minimum of H); it follows that s is of constant sign. Hence, the functional dependence of F on H is monotonic. The function F in (A.14) is therefore equivalent to H, *as stated in* Theorem I.

APPENDIX B

Mathematics of Cross-Entropy Minimization

We derive the general solution for cross-entropy minimization given arbitrary constraints, and we illustrate the result with the important cases of exponential and Gaussian densities. In general, however, it is difficult or impossible to obtain a closed-form, analytic solution expressed directly in terms of the known expected values rather than in terms of the Lagrangian multipliers. We therefore discuss numerical techniques for obtaining the solution, namely the Newton-Raphson method. This method is the basis for a computer program that solves for the minimum cross-entropy posterior given an arbitrary prior and arbitrary expected-value constraints.

Given a positive prior density p and equality constraints

$$\int q(\underset{\sim}{x})\, d\underset{\sim}{x} = 1 \,, \tag{B.1}$$

$$\int f_k(\underset{\sim}{x}) q(\underset{\sim}{x})\, d\underset{\sim}{x} = \bar{f}_k \,, \qquad (k = 1, \ldots, m) \,, \tag{B.2}$$

the standard method for seeking an extremum of

$$H(q,p) = \int q(\underset{\sim}{x}) \log \frac{q(\underset{\sim}{x})}{p(\underset{\sim}{x})}\, d\underset{\sim}{x} \,,$$

subject to the constraints, is to introduce Lagrangian multipliers λ_0 and λ_k $(k = 1, \ldots, m)$ corresponding to the constraints, forming the expression

$$\int q(\underset{\sim}{x}) \log \frac{q(\underset{\sim}{x})}{p(\underset{\sim}{x})}\, d\underset{\sim}{x} + \lambda_0 \int q(\underset{\sim}{x})\, d\underset{\sim}{x} + \sum_{k=1}^{m} \lambda_k \int f_k(\underset{\sim}{x}) q(\underset{\sim}{x})\, d\underset{\sim}{x} \,,$$

and to equate the variation, with respect to q, of this quantity to zero:

$$\log \frac{q(\underset{\sim}{x})}{p(\underset{\sim}{x})} + 1 + \lambda_0 + \sum_{k=1}^{m} \lambda_k f_k(\underset{\sim}{x}) = 0 \ .$$

Solving for q leads to

$$q(\underset{\sim}{x}) = p(\underset{\sim}{x}) \exp\left(- \lambda_0 - 1 - \sum_{k=1}^{m} \lambda_k f_k(\underset{\sim}{x}) \right) . \tag{B.3}$$

It is necessary to choose λ_0 and the λ_k so that the constraints are satisfied. In the presence of the constraint (B.1) we may rewrite the remaining constraints in the form

$$\int f_k(\underset{\sim}{x}) q(\underset{\sim}{x}) \, dx = 0 \tag{B.4}$$

by redefining the f_k: write $f_k(\underset{\sim}{x})$ for what was previously written as $f_k(\underset{\sim}{x}) - \bar{f}_k$. Now, if we find values for the λ_k such that

$$\int f_i(\underset{\sim}{x}) p(\underset{\sim}{x}) \exp\left(- \sum_{k=1}^{m} \lambda_k f_k(\underset{\sim}{x}) \right) d\underset{\sim}{x} = 0 \quad ,$$

$$(i = 1, \ \ldots \ , m) \ , \tag{B.5}$$

we are assured of satisfying (B.4); and we can then satisfy (B.1) by setting

$$\lambda_0 = - 1 + \log \int p(\underset{\sim}{x}) \exp\left(- \sum_{k=1}^{m} \lambda_k f_k(\underset{\sim}{x}) \right) d\underset{\sim}{x} \ .$$

The situation for inequality constraints is only slightly more complicated. Suppose we replace all the equal signs in (B.2) by $\leqslant$. (We lose no generality thereby: we can change inequalities with $\geqslant$ into inequalities

with $\lesseqgtr$ by changing the signs of the corresponding f_k and $\bar{f}_k$, and any equality constraint is equivalent to a pair of inequality constraints.) The q that minimizes H(q,p) subject to the resulting constraints will in general satisfy equality for certain values of k in the modified (B.2), while strict inequality will hold for the rest. We can still use the solution (B.3), subjecting the Lagrange multipliers to the conditions $\lambda_k \lesseqgtr 0$ for k such that equality holds in the constraint, and $\lambda_k = 0$ for k such that strict inequality holds in the constraint.

It unfortunately is usually impossible to solve (B.5) for the λ_k explicitly, in closed form; however, it is possible in certain important special cases. For example, consider the case in which the prior $p(\underset{\sim}{x})$ is a multivariate exponential,

$$p(\underset{\sim}{x}) = \prod_{k=1}^{n} (1/a_k) \exp(x_k/a_k) ,$$

where $\underset{\sim}{x} = (x_1,\ldots,x_n)$ and the x_k each range over the positive real line, and in which the constraints are

$$\int d\underset{\sim}{x}\; x_k q(\underset{\sim}{x}) = \bar{x}_k , \qquad \text{(B.6)}$$

k = 1, ..., n. Solving (B.5) in order to express the minimum cross-entropy posterior directly in terms of the known expected values $\bar{x}_k$ yields

$$q(\underset{\sim}{x}) = \prod_{k} (1/\bar{x}_k) \exp(x_k/\bar{x}_k).$$

Thus, the density remains multivariate exponential, with the prior mean values a_k being replaced by the newly learned values $\bar{x}_k$.

Now consider the case in which the x_k range over the entire real line, and in which the prior density is Gaussian,

$$p(\underset{\sim}{x}) = \prod_k (2\pi b_k)^{-1/2} \exp\left[(x_k - \bar{x}_k)^2/2b_k\right] .$$

Suppose that the constraints are (B.6) and

$$\int d\underset{\sim}{x}\, (x_k - \bar{x}_k)^2 q(\underset{\sim}{x}) = v_k .$$

In this case the minimum cross-entropy posterior is

$$q(\underset{\sim}{x}) = \prod_k (2\pi v_k)^{-1/2} \exp\left[(x_k - \bar{x}_k)^2/2v_k\right] .$$

Thus, the density remains multivariate Gaussian, with the prior means and variances being replaced by the newly learned values.

Here is an example of a simple problem for which the solution of (B.5) cannot be expressed in closed form. Consider a discrete system with n states x_j and prior probabilities $p(x_j) = p_j$ $(j = 1, \ldots, n)$. The discrete form of (B.1) is

$$\sum_{j=1}^{n} q_j = 1 , \tag{B.7}$$

where $q_j = q(x_j)$. Suppose the only other constraint is that the mean m of the indices j is prescribed: $f(x_j) = j$, and

$$\sum_{j=1}^{n} j q_j = m . \tag{B.8}$$

Equation (B.3) becomes

$$q_j = p_j \exp(-\lambda_0 - 1 - j\lambda) ,$$

which we write as

$$q_j = ap_j z^j$$

by introducing the abbreviations

$$a = \exp(-\lambda_0 - 1) , \qquad z = \exp(-\lambda) .$$

From (B.7) and (B.8) we then obtain

$$a = \left(\sum_{j=1}^{n} p_j z^j \right)^{-1}$$

and

$$\sum_{j=1}^{n} (j-m) p_j z^j = 0 . \tag{B.9}$$

The problem then reduces to finding a positive root of the polynomial in (B.9). As in the continuous case, there are special forms for the prior that lead to important particular solutions. But when $n > 5$, the roots of the polynomial (other than zero) cannot in general be written as explicit, closed-form expressions in the coefficients for arbitrary priors. Numerical methods of solution therefore become important. Our obtaining a polynomial equation in the present example was an accidental consequence of the fact that the values of the constraint function f formed a subset of an arithmetic progression ($j = 1, 2, \ldots$). Thus, for more general types of problems, numerical methods are even more important.

One such method is the Newton-Raphson method, which is for finding solutions for systems of equations that, like (B.5), are of the form

$$F_i(\lambda_1, \ldots, \lambda_m) = 0 , \qquad (i = 1, \ldots, m) . \tag{B.10}$$

The method starts with an initial guess at the solution, $\underset{\sim}{\lambda}^{(1)} = (\lambda_1^{(1)}, \ldots, \lambda_m^{(1)})$, and produces further approximate solutions $\underset{\sim}{\lambda}^{(2)}$, $\underset{\sim}{\lambda}^{(3)}$, ... in succession. If the initial guess $\underset{\sim}{\lambda}^{(1)}$ is close enough to a solution of (B.10), if the F_i are continuously differentiable, and if the Jacobian $[\partial F_i/\partial \lambda_j]$ is nonsingular, then the $\underset{\sim}{\lambda}^{(r)}$ will converge to the solution in the limit as $r \to \infty$.

The method is based on the fact that, for small changes $\underset{\sim\sim}{\Delta\lambda}^{(r)}$ in the arguments $\underset{\sim}{\lambda}^{(r)}$, we have the approximate equality

$$F_i(\underset{\sim}{\lambda}^{(r)} + \underset{\sim\sim}{\Delta\lambda}^{(r)}) \approx F_i(\underset{\sim}{\lambda}^{(r)}) + \sum_{k=1}^{m} \frac{\partial F_i(\underset{\sim}{\lambda}^{(r)})}{\partial \lambda_k^{(r)}} \Delta\lambda_k^{(r)}$$

up to a term of order $o(\underset{\sim\sim}{\Delta\lambda}^{(r)})$. We therefore take $\underset{\sim\sim}{\Delta\lambda}^{(r)}$ to be a solution of the linear equation

$$\sum_{k=1}^{m} \frac{\partial F_i(\underset{\sim}{\lambda}^{(r)})}{\partial \lambda_k^{(r)}} \Delta\lambda_k^{(r)} = -F_i(\underset{\sim}{\lambda}^{(r)}) \tag{B.11}$$

and set

$$\underset{\sim}{\lambda}^{(r+1)} = \underset{\sim}{\lambda}^{(r)} + \underset{\sim\sim}{\Delta\lambda}^{(r)} .$$

When F_i is given by the discrete form of the left-hand side of (B.5), we have

$$F_i(\underset{\sim}{\lambda}^{(r)}) = \sum_{j=1}^{n} f_{ij} p_j \exp\left(- \sum_{u=1}^{m} \lambda_u^{(r)} f_{uj} \right), \tag{B.12}$$

$$\frac{\partial F_i(\underset{\sim}{\lambda}^{(r)})}{\partial \lambda_k} = - \sum_{j=1}^{n} f_{ij} f_{kj} p_j \exp\left(- \sum_{u=1}^{m} \lambda_u^{(r)} f_{uj} \right), \tag{B.13}$$

where $f_{ij} = f_i(x_j)$. With the abbreviation

$$g_j = p_j^{1/2} \exp\left(- \frac{1}{2} \sum_{u=1}^{m} \lambda_u^{(r)} f_{uj} \right),$$

we express the right-hand sides of (B.12) and (B.13) in matrix notation as

$$(\underset{\sim}{f}\, \mathrm{diag}(\underset{\sim}{g})\, \underset{\sim}{g})_i ,$$

$$(\underset{\sim}{f}\, \mathrm{diag}(\underset{\sim}{g})^2\, \underset{\sim}{f}^t)_{ik} ,$$

where $\mathrm{diag}(\underset{\sim}{g})$ is the diagonal matrix whose diagonal elements are the g_j, and $\underset{\sim}{f}^t$ is the transpose of $\underset{\sim}{f}$. The solution of (B.11) is then given by

$$\underset{\sim\sim}{\Delta\lambda}^{(r)} = \left[(\underset{\sim}{f}\, \mathrm{diag}(\underset{\sim}{g})^2\, \underset{\sim}{f}^t)^{-1}\, \underset{\sim}{f}\, \mathrm{diag}(\underset{\sim}{g}) \right] \underset{\sim}{g} .$$

We remark that the quantity in brackets is the Moore-Penrose generalized inverse [58] of the matrix $\underset{\sim}{f}\, \mathrm{diag}(\underset{\sim}{g})$. The approach just described has been made the basis for a computer program [59], written in APL, for solving cross-entropy minimization problems with arbitrary positive discrete priors $\underset{\sim}{p}$ and equality constraints specified by matrices $\underset{\sim}{f}$. The approach is particularly convenient for programming in APL since the generalized inverse is a built-in APL primitive function [60]. To solve a minimum-cross-entropy problem with 500 states and 10 constraints, the program typically requires 15 seconds of CPU time when running under the APL SF interpreter on a DEC-10 system with a KI central processor.

ACKNOWLEDGMENTS

The authors thank A. Ephremides and W. S. Ament for their reviews of an earlier version of this paper.

REFERENCES

1. E. T. Jaynes, "Information Theory and Statistical Mechanics I," Phys. Rev. 106, 1957, pp. 620-630.

2. E. T. Jaynes, "Information Theory and Statistical Mechanics II," Phys. Rev. 108, 1957, pp. 171-190.

3. E. T. Jaynes, "Information Theory and Statistical Mechanics," in Statistical Physics, V. 3 (Brandeis Lectures), (K. W. Ford, Ed.), Benjamin, Inc., New York, 1963, pp. 182-218.

4. E. T. Jaynes, "Foundations of Probability Theory and Statistical Mechanics," in M. Bunge, Ed., Delaware Seminar in the Foundations of Science, Vol. I, Springer-Verlag, New York, 1967, pp. 77-101.

5. O. C. de Beauregard and M. Tribus, "Information Theory and Thermodynamics," Helvetica Physica Acta 47, 1974, pp. 238-247.

6. M. Tribus, Thermostatics and Thermodynamics, D. Van Nostrand, Princeton, New Jersey, 1961.

7. A. Katz, Principles of Statistical Mechanics - The Information Theory Approach, W. H. Freeman Company, New York, 1967.

8. A. Hobson, Concepts in Statistical Mechanics, Gordon and Breach, New York, 1971.

9. I. J. Good, "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables," Annals Math. Stat. 34, 1963, pp. 911-934.

10. J. P. Noonan, N. S. Tzannes, and T. Costello, "On the Inverse Problem of Entropy Maximizations," IEEE Trans. Inform. Theory IT-22, 1976, pp. 120-123.

11. M. Tribus, Rational Descriptions, Decisions, and Designs, Pergamon Press, New York, 1969.

12. M. Tribus, "The Use of the Maximum Entropy Estimate in the Estimation of Reliability," in Recent Developments in Information and Decision Processes, (R. E. Machol, and D. Gray, Eds.), Macmillan, New York, 1962, pp. 102-139.

13. V. E. Benes, Mathematical Theory of Connecting Networks and Telephone Traffic, Academic Press, New York, 1965.

14. A. E. Ferdinand, "A Statistical Mechanics Approach to Systems Analysis," IBM J. Res. Develop., 1970, pp. 539-547.

15. J. E. Shore, "Derivation of Equilibrium and Time-Dependent Solutions to M/M/oo//N and M/M/oo Queueing Systems Using Entropy Maximization, Proceedings 1978 National Computer Conference, AFIPS, 1978, pp. 483-487.

16. M. Chan, "System Simulation and Maximum Entropy," Operations Research 19, 1971, pp. 1751-1753.

17. E. T. Jaynes, "New Engineering Applications of Information Theory," in Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability, (J. L. Bogdanoff, Ed.), John Wiley, New York, 1963, pp. 163-203.

18. M. Tribus and G. Fitts, "The Widget Problem Revisited," IEEE Trans. on Systems Science and Cybernetics SSC-4, 1968, pp. 241-248.

19. B. Ramakrishna Rau, "The Exact Analysis of Models of Program Reference Strings," Stanford Electronics Laboratory Technical Report 124, (SU-SEL-77-003), Stanford University, December, 1976.

20. A. E. Ferdinand, "A Theory of General Complexity," Int. J. General Syst. 1, 1974, pp. 19-33.

21. M. Takatsuji, "An Information-Theoretical Approach to a System of Interacting Elements," Biol. Cybernetics 17, 1975, pp. 207-210.

22. J. M. Cozzolino and M. J. Zahner, "The Maximum-Entropy Distribution of the Future Market Price of a Stock," Operations Research 21, 1973, pp. 1200-1211.

23. E. T. Jaynes, "Probability Theory in Science and Engineering," Field Research Laboratory, Socony Mobil Oil Company, Inc., Colloquium Lectures in Pure and Applied Science No. 4, 1958.

24. E. T. Jaynes, Probability Theory in Science and Engineering, unpublished lecture notes (available from Physics Department, Washington Univ., St. Louis, Mo.), 1972.

25. E. T. Jaynes, "Prior Probabilities," IEEE Trans. on Systems Science and Cybernetics SSC-4, 1968, pp. 227-241.

26. J. Burg, "Maximum Entropy Spectral Analysis," Ph.D. Dissertation, Stanford University, 1975 (University Microfilms No. 75-25,499).

27. J. G. Ables, "Maximum Entropy Spectral Analysis," Astron. Astrophys. Suppl. 15, 1974, pp. 383-393.

28. T. J. Ulrych and T. N. Bishop, "Maximum Entropy Spectral Analysis and Autoregressive Decomposition," Reviews of Geophysics and Space Physics 43, No. 1, 1975, pp. 183-200.

29. S. J. Wenecke and L. R. D'Addario, "Maximum Entropy Image Reconstruction," IEEE Trans. on Computers C-26, 1977, pp. 351-364.

30. I. J. Good, Probability and the Weighing of Evidence, Charles Griffen, London, 1950.

31. S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.

32. A. Wehrl, "Entropy," Rev. Mod. Phys. 50, No. 2, 1978, pp. 221-260.

33. A. Hobson and B. Cheng, "A Comparison of the Shannon and Kullback Information Measures," J. Stat. Phys. 7, No. 4, 1973, pp. 301-310.

34. R. Johnson, "Axiomatic Characterization of the Directed Divergences and Their Linear Combinations," submitted to IEEE Trans. Inf. Theory.

35. R. S. Ingarden and A. Kossakowski, "The Poisson Probability Distribution and Information Therodynamics," Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys. 9, No. 1, 1971, pp. 83-85.

36. G. S. Arnold and J. L. Kinsey, "Information Theory for Marginal Distributions: Application to Energy Disposal in an Exothermic Reaction," J. Chem. Phys. 67, No. 8, 1977, pp. 3530-3532.

37. R. L. Kashyap, "Prior Probability and Uncertainty," IEEE Trans. Inf. Theory IT-17, 1971, pp. 641-650.

38. R. Johnson, "Comments on 'Prior Probability and Uncertainty'", IEEE Trans. Inf. Theory, to be published.

39. P. M. Lewis II., "Approximating Probability Distributions to Reduce Storage Requirements," Inf. and Control 2, 1959, pp. 214-225.

40. C. E. Shannon, "A Mathematical Theory of Communication," Bell System Tech. Jour., 27, 1948, pp. 379-423.

41. C. E. Shannon, "A Mathematical Theory of Communication," Bell System Tech. Jour., 27, 1948, pp. 623-656.

42. J. S. Rowlinson, "Probability, Information, and Entropy," Nature 225, 28 March 1970, pp. 1196-1198.

43. J. M. Jauch and J. G. Baron, "Entropy, Information, and Szilard's Paradox," Helvetica Physics Acta 45, 1972, p. 220.

44. K. Friedman and A. Shimony, "Jaynes' Maximum Entropy Prescription and Probability Theory," J. Stat. Phys. 3, No. 4, 1971, pp. 381-384.

45. M. Tribus and H. Motroni, "Comments on the Paper Jaynes' Maximum Entropy Prescription and Probability Theory," J. Stat. Phys. 4, No. 2/3, 1972, pp. 227-228.

46. A. Hobson, "The Interpretation of Inductive Probabilities," J. Stat. Phys. 6, No. 2/3, 1972, pp. 189-193.

47. A. I. Khinchin, Mathematical Foundations of Information Theory, Dover, New York, 1957.

48. A. Hobson, "A New Theorem of Information Theory," J. Stat. Phys. 1, No. 3, 1969, pp. 383-391.

49. Pl. Kannappan, "On Shannon's entropy, directed divergence, and inaccuracy," Z. Wahrscheinlichkeitstheorie verw. Geb., Vol. 22, pp. 95-100, 1972.

50. Pl. Kannappan, "On directed divergence and inaccuracy," Z. Wahrscheinlichkeitstheorie verw. Geb., Vol. 25, pp. 49-55, 1972.

51. J. Aczel and Z. Daroczy, On Measures of Information and Their Characterizations, Academic Press, N. Y. 1975.

52. R. T. Cox, "Probability, Frequency, and Reasonable Expectation," Am. J. Phys. 14, 1946, pp. 1-13.

53. R. T. Cox, The Algebra of Probable Inference, John Hopkins Press, Baltimore, 1961.

54. L. Janossy, "Remarks on the Foundation of Probability Calculus," Acta Phys. Acad. Hungar. 4, 1955, pp. 333-349.

55. J. Aczel, "A Solution of Some Problems of K. Borsuk and L. Janossy," Acta Phys. Acad. Hungar. 4, 1955, pp. 351-362.

56. J. Aczel, Lectures on Functional Equations and their Applications, Academic Press, N. Y., 1966.

57. C. T. Ng, "Representation for Measures of Information With the Branching Property," Inf. and Control 25, 1974, pp. 45-56.

58. A. E. Albert, Regression and the Moore-Penrose Pseudoinverse, Academic Press, N. Y. 1972.

59. R. W. Johnson, "Computer Programs for Determining Probability Distributions by the Principles of Maximum Entropy and Minimum Cross-Entropy," in preparation.

60. M. A. Jenkins, "The Solution of Linear Systems of Equations and Linear Least Squares Problems in APL," IBM, N. Y., Scientific Center Technical Report Number 320-2989, June, 1970.

*U.S. GOVERNMENT PRINTING OFFICE: 1978 281-484/189 1-3